Identifying the urban communities of New York City using bikeshare data from NYC CitiBike

Colin Broderick - colin@cbroderick.me
26th February 2015

Supervisors
Mark Padgham, PhD. (WWU)
Marco Painho, PhD. (UNL)
Oscar Belmonte, PhD. (UJI)

Dissertation submitted 26th February 2015 to Institut für Geoinformatik, Westfälische Wilhelms-Universität Münster (WWU), Germany in partial fulfillment of requirements for Degree of Master of Science in Geospatial Technologies.

Institutions
Institut für Geoinformatik, Westfälische Wilhelms-Universität Münster (WWU), Germany.
NOVA Information Management School, Universidade Nova de Lisboa (UNL), Portugal.
Universitat Jaume I (UJI), Castellón, Dept. Lenguajes y Sistemas Informaticos, Castellón, Spain

“Identifying the urban communities of New York City using bikeshare data from NYC CitiBike”

Colin Broderick - colin@cbroderick.me
26th February 2015

Supervisors
Mark Padgham, PhD. (WWU)
Marco Painho, PhD. (UNL)
Oscar Belmonte, PhD. (UJI)

Dissertation submitted 26th February 2015 to Institut für Geoinformatik, Westfälische Wilhelms-Universität Münster (WWU), Germany in partial fulfillment of requirements for Degree of *Master of Science in Geospatial Technologies*.

Institutions
Institut für Geoinformatik, Westfälische Wilhelms-Universität Münster (WWU), Germany.
NOVA Information Management School, Universidade Nova de Lisboa (UNL), Portugal.
Universitat Jaume I (UJI), Castellón, Dept. Lenguajes y Sistemas Informaticos, Castellón, Spain

Date	Version	Distributed To
22/01/2015	PRELIM. DRAFT	MP, MaP, OB, IG, EP
26/02/2015	FINAL VERSION	KH

Declaration

I, CB ___________________________, hereby declare that I have written this thesis independently, unless where clearly stated otherwise. I have used only the sources, the data and the support that I have clearly mentioned. This thesis has not been submitted for conferral of degree elsewhere.

Signature Colin Broderick

Muenster,

February 26, 2015

Acknowledgement

I would like to thank everyone who helped me reach this final stage, you too deserve a pat on the back for putting up with my half finished drafts and sentences that go nowhere.

I would like to especially thank my parents, Yasmin Hamed, Ciaran Staunton and my supervisor, Mark Padgham for your inspiration, support and above all else your endless patience as I slowly pieced this work together.

Last but not least thank you to all of the European Union taxpayers and your education tax contributions, for your financial support which allowed me to undertake this Masters and dissertation.

Abstract

This study identifies a measure of the cultural importance of an area within a city. It does so by making use of origin-destination trip data and the bike stations of the bike share system in New York City as a proxy to study the city. Rarely is movement in the city studied at such a small scale. The change in strength of the similarity of movement between each station is studied. It is the first study to provide this measure of importance for every point in the system. This measure is then related to the characteristics which make for vibrant city communities, namely highly mixed land use types. It reveals that the spatial pattern of important areas remains constant over differing time periods. Communities are then characterised by the land uses surrounding these stations with high measures of importance. Finally it identifies the areas of global cultural importance alongside the areas of local importance to the city.

Keywords

GIS

Cities

Open Data

City Study

R

Python

Bikeshare Schemes

Walkability

Urban Planning

Mapping

Geospatial

Urban Community Hubs

Table of Contents

Figures

Figure 19 - New York K-Values on Weekdays (Monday - Friday) with high k-value communities identified

Figure 20 - New York K-Values on Weekends (Saturday - Sunday) with high k-value communities identified

Figure 21 - New York total trips from K-Values plotted against PLUTO land use types.

Tables

Acronyms

CRAN	Comprehensive R Archive Network
CSV	Comma Separated Value
EPSG	European Petroleum Survey Group
HTML	HyperText Markup Language
JSON	JavaScript Object Notation
NLS	Non-linear Least Squared model
Numpy	Numerical Python
NYC	New York City
OSM	OpenStreetMap
Pandas	Python Data Analysis Library
PLUTO	Property Land Use Tax lot Output
SQL	Standard Query Language
SSR	Sum of Squared Residuals
XML	Extensible Markup Language
UTM	Universal Transverse Mercator
WGS	World Geodetic System

1. Introduction

The spatial structure of cities is changing and as such, transport has to adapt to this new structure. In the past it had been argued (Park et al. 1925) that cities emulated an ecosystem, similar to those studied by ecologists, in that, like plants, they would grow from the centre outwards in graduated rings. The oldest and least desirable parts of the city were at its core, while the most prosperous areas were around the edges. This theory is known as Concentric Zone Theory which was developed by the father of modern planning Robert E. Park. (Park et al. 1925)

In this post-Parkian era we have seen that cities are no longer of the monocentric form in which employment is located in one place and in which the majority of commuting trips radiate from this location. Due to the change in economic practices from heavy industry to that of the services industry, it is now possible to live and work in the same area. Cities have moved from this monocentric structure to a more polycentric structure in which commuting is much more evenly distributed making it more challenging to serve these areas with public transport. Many trips now comprise short direct distances to access services and work. (Anas et al., 1998; Kloosterman and Musterd, 2001)

We are now on the cusp of the age of the so-called “smart city”, which is characterised by low-carbon and low-pollution transportation, integration of sensing technologies, shared societal resources, and significant benefits to public health. (Midgley, 2009) Bikeshare systems are rapidly becoming a new-old solution to mass transportation within the close confines of city centres. At present there are over 450 of these systems in operation worldwide (Meddin and DeMaio, 2014) including implementations in New York, London, and my home city of Dublin. These systems comprise of bike stations which include a number of stands where a user can check in and check out a bike as part of their journey.

These systems have been described previously as networks which display distinct communities which were identified by the spatio-temporal characteristics of all journeys within the system. Clustering methods were used to identify these communities. (Austwick et al., 2013; Lathia et al., 2012).

Padgham (2014, unpublished) proposes a theory on how urban spaces become partitioned into distinct clusters by using hierarchical clustering techniques to derive this structure using bike share data from London.

It is intended that this thesis will build upon the methodologies of both Pagham (2012 and 2014) and Austwick et al. (2013) to identify the communities of New York City. New York City has a bike system comprising of 332 bike stations and users have made over fifteen million trips since the system opened on 27th May 2013. (CitiBikeNYC, 2014). All models of monocentric cities assume a single measure of the similar of movement which radiates to and from a single centre. This thesis measures this similarity of movement for every monocentric center of the city represented by the stations of the CitiBike scheme.

This research makes best use of both open source data and open source software. All of the analysis and preparatory code is available for reuse openly through github[1], the links to which are available in Appendix 1.

As such, this dissertation takes the following structure:

- Section 8 describes the theory and models used;
- Section 9 details the data sources used;
- Section 10 contains the methods, written in a first-person user manual type style in the spirit of open and reproducible research;
- Section 11 contains both results and discussion; and
- Followed finally by conclusions at Section 12.

[1] See - https://github.com/rustyb/bike-correlations/tree/master

2. Research Objectives

The role of the spatial planner is to attempt to control space and how that space is used especially within urban areas. In the past various models have been proposed to describe how urban areas develop. These models have been implemented without little thought for testing their underlying assumptions. This is particularly true when it comes to assumptions based around the movement of people throughout urban areas. This research aims in part to help measure these movements at a fine scale so that in the future these models can be tested. The following are the research objectives of this research:

1. To Identify a measure of cultural importance for each point in transport system.
2. To develop a method which can be used and repeated on other cities which uses open source tools and data sources.
3. To arrive at an understanding of transit behaviour as a cultural and community aspect within a City.
4. To arrive at an understanding of how this transit behaviour is related to the mix of land use types within a City.

3. Theory

The principal aim of this research is to study the relationship between bike stations in the New York City using origin-destination data. Statistical Analysis provides certain measures which allow us to study the relationships between things and these are called correlations. A correlation is a single number which describes the relationship between two or more things.

Examining this mutual relationship between two or more things has allowed the study of many types of spatial phenomena such as the spread of disease throughout a population (Bolker et al. 1996), the relationship between the the types of crime committed and the dominant land use in an area (Lockwood, D. 2007) and the study of land use and transport dynamics within cities (Batty, M. 2013).

The bike stations which make up the New York CitiBike scheme and the trips made between each of these stations will be used to measure the strength of the relationship between different parts of New York City.

These bike stations act as the nodes which anchor movement within the city. It is generally assumed that human movement, especially in cities, is due to the desire to move from one origin to a destination (Jukka-Pekka, et al., 2011). The 332 bike stations act as these origins and destinations. All movements begin and end at one of these stations.

3.1. Correlations

One way to study the similarities of each station to the others is to construct/calculate a correlation from the number of trips made to and from that station to another, and to repeat for each and every other station pair.

To explain this process, it is helpful firstly to examine a simple example: the stations labelled 1 and 2 in Figure 1 below. These two stations while nearby in space may not be similar to each other in any other way. The red points represent stations between which correlations will be made. Thicker lines represent higher volumes of trips. Pairwise correlations between each red point are the correlation between the thickness of all black lines and corresponding dashed grey lines. The correlation is a the direct measure of the overlap in variance of trip numbers between the two stations.

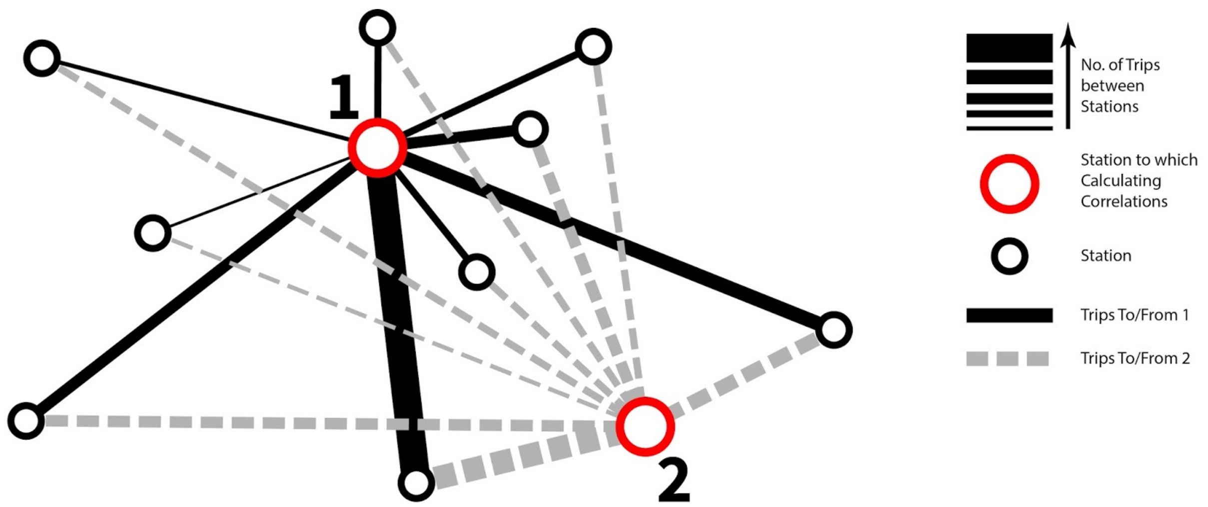


Figure 1 - How interstation correlations are calculated.

Within other scientific disciplines which study the dynamics of movement such as ecology, engineering and economics, it has been shown that this relationship between things generally decreases with distance. (Nekola et al., 1999)

According to Tobler's first law of geography *“everything is related to everything else, but near things are more related than distant things”*. (Tobler, 1970) Therefore, these correlations between stations should be related to distance in some way. The following distributions are likely to be expected where the correlation generally decreases with distance:

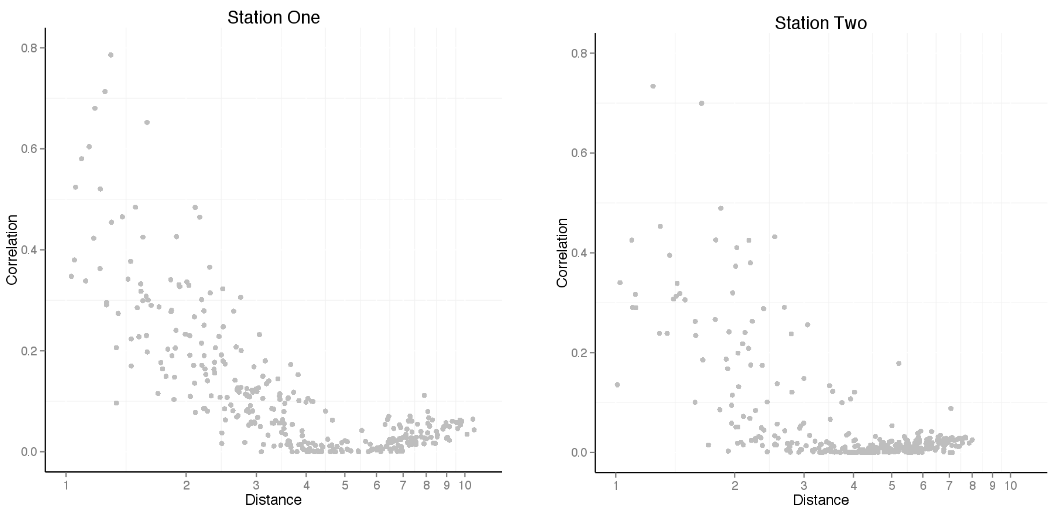


Figure 2 - Example distribution of correlations and distances for Station One and Station Two.

3.2. Distance Decay Models

The principal aim of this research is to study the spatial structure of this distance decay and relate it to the current structure of the city. Given this aim, a number of distance decay models must be fitted to the distributions in order to ascertain the best decay model.

A non-linear least squared model (NLS) will be used to fit a function to the station distance and correlation data. NLS is a *“mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets (‘the residuals’) of the points from the curve”.* (Weisstein, 2014)

There are various distance decay functions which take the form of NLS models. The most common types are, exponential growth/decay (Smyth, 2002), Gaussian , Inverse, and

Gravity (Power Law). Each of these models determines the weight to attribute to distance in a different way. This is best explained visually with a diagram, see Figure 3 below. As can be seen, the Gaussian takes the form of an “S” shaped curve. The difference is that in the Gravity model distance is squared which results in a curve which decays much faster with distance.

Economic Gravity based decay models are traditionally used when examining the spatial relationships at play in cities, such as the segmentation in urban housing markets (Schnare and Raymond, 1976), commuting patterns and hospital patients (Cheng and Howard, 1999). Each of the four models discussed above were tested on the data from two sample stations from the New York CitiBike data. The results of these tests determined the best model to use for fitting to the other stations.

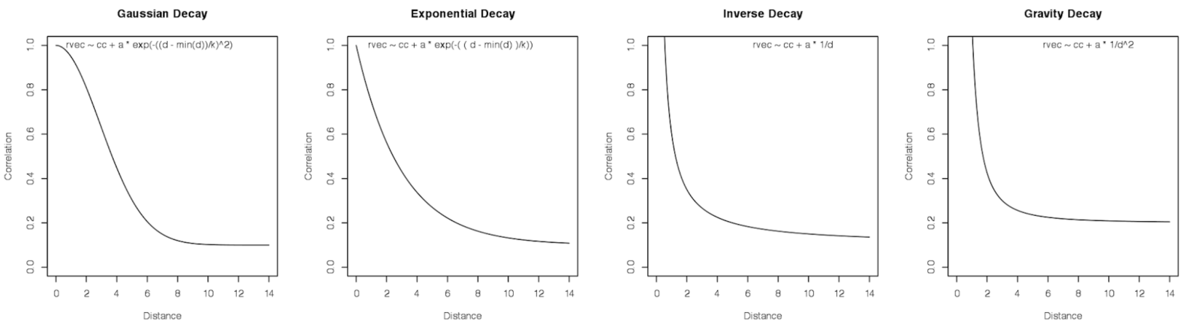


Figure 3 - Distance decay model examples.

The statistical measure, the sum of squared residuals (SSR), is used to quantitatively compare each model. This is a measure of the variance of the real data and grey points to that of the fitted values of the model. These fitted values are represented by the respective line. The lower this value is the better fit the model is to the data as the squares are minimized.

Padgham (2012) showed that the distribution of trips and distance follows a Gaussian distribution. From here on it assumed that the Gaussian distance decay model is the best fit for the resulting station correlations and inter station distances.

The Gaussian function is fitted to the correlations between each station and the shortest distance between them as follows:

R^2 ~ interstation distance

The formula for the function is as follows:

$$R^2 \sim e\text{-}(d/k)^2$$

Where:

- **R^2** is the correlation between each station pair, for example station 1 and every other station.
- **d** is the distance between that station and every other.
- **k** is the width of Gaussian decay.

The focal point of this whole body of research is to derive the K-value for each station. High k-values describe stations at which movement to/from is similar to the movement to/from nearby stations. These nearby stations will also have similarly elevated K-values. The k-value quantifies the spatial range at which movement towards and away from each point is similar or highly coordinated throughout the entire system. Thus high k-value stations can be interpreted as having a large scale importance whereas stations with low K-values are relatively less important.

Using NLS to model this function is an iterative process and as such, the model's parameters will be adjusted until convergence is achieved; when the squares have been minimized.

This function is fitted for every station in the network, so that a K-value can be derived for every single station. This is represented graphically below, Figure 4. The left side shows the fitted Gaussian and the right side, different derived Gaussians for other stations. All station correlations are positive, and as such each ranges between 0 and 1.

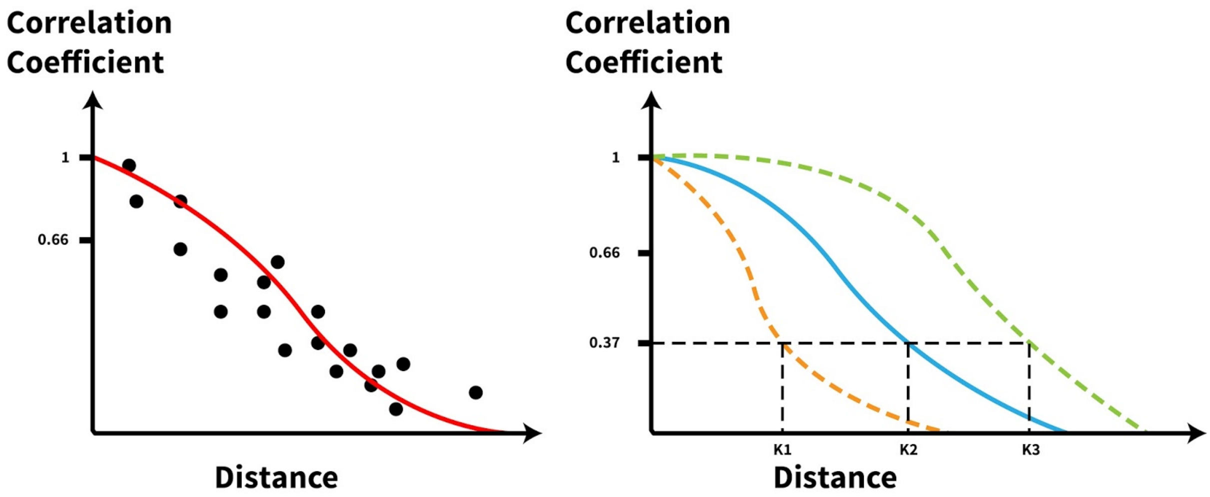


Figure 4 - Gaussian fitted to distribution of correlations and distances from an example station. (Left) Gaussians fitted to a number of stations and their respective K-values indicated. (Right)

These resulting K-values can then be used to describe the strength of the relationship of one station when compared to its surrounding stations.

3.3. What does the K-value mean?

The results can then be overlaid on the New York City in order to identify any prevailing spatial patterns in the resulting K-values. Consider Figure 5 where the size of each point represents the magnitude of the K-value.

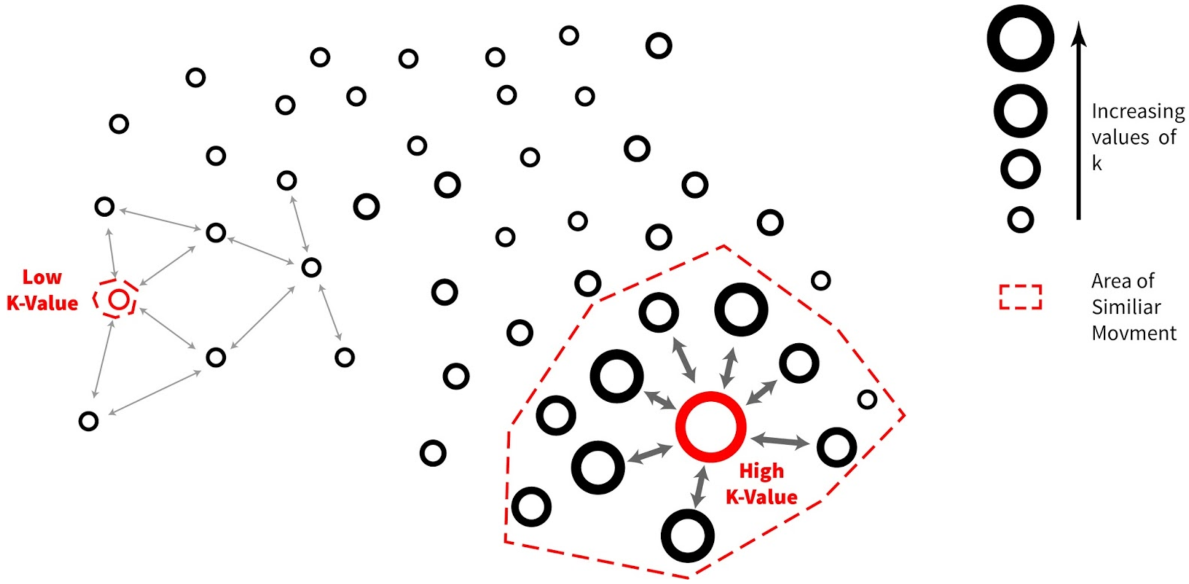


Figure 5 - What high K-values and low K-values signify.

As can be seen there are a number of stations which show high K-values, and these generally are nearby each other. Meanwhile, as one moves farther from the area of high K-values, the station K-values decrease. The stations that have little broad-scale importance and/or no relationship with surrounding points have low K-values.

Figure 5 illustrates the phenomenon which arises in areas of high K-values. The stations nearby the high K-value stations will be clustered together with nearby stations of similar K-values. It is understood that people often orientate themselves towards places of significance within transport systems, and their movements can be directed both to/from these places (Padgham, 2012).

Movement from anywhere in the city towards/away from a hotspot of high K-value stations remains very similar in regard to all stations in the vicinity of that hotspot. A high K-value may thus be interpreted as reflecting the global-scale cultural or geographical, importance of the station. Conversely a low K-value should be interpreted as a station which is of local importance only and not strongly related to the other stations within its vicinity.

It is widely known that large cities are not simply made up of a single urban conurbation with one centre to which everybody moves to and from (Kloosterman and Musterd, 2001). The area within a city which people are from or live is central to the cultural desirability of destinations they may choose. Whether they are morphologically or functionally polycentric as observed by Berger and Meijers (2012), certain areas within a city have better/diverse services compared to others, or specialize in particular things which cannot be found in other areas. This leads to promoting higher levels of desirability of one area over another. The urban community hubs of a city are the places where people live, work, study, leisure. They provide the essential functions to sustain people with work and resources.

There are certain ‘intangible’ considerations necessary to fully understand the desirability within cities which come into play (van Lenthe et al., 2004). One is that travelling on less managed and more discrete modes like walking and cycling (compared to bus, train and car) are open to flexibility, and less predictability and cultural desirability (Banister, 2008). This means that the destination end point of such modes cannot be a precise location proxy for the desirability of the destination in the same way as bike use, which requires smaller less obtrusive infrastructure. There is also more spontaneity and trips are open to less fixed cultural desirability when compared to other modes (Mokhtarian et al, 2006).

Evidence from Saelens et al. (2003) shows that neighbourhoods of *“higher density, greater connectivity, and diverse land use mix report higher rates of walking/cycling for utilitarian purposes than those of low-density, poorly connected, and single land use neighborhoods”*.

These are the areas of high desirability, which are characterised by having high K-values. Given this phenomenon urban community hubs can then be identified as the areas in the vicinity of high k-values as proposed in the theory section. Land use patterns within these areas would be expected to be classified as predominantly mixed use.

3.4. Urban Community Hubs

City planning emerged as a profession primarily as a response to the polluted inner cities filled with the vast polluting factories spawned by the industrialisation in 1920s (Brantz and Dümpelmann, 2011). It is clear that there is a link between the layout and uses of our city and human health. Planners set about separating these polluting and damaging industries from the places people lived (Taylor, 1998).

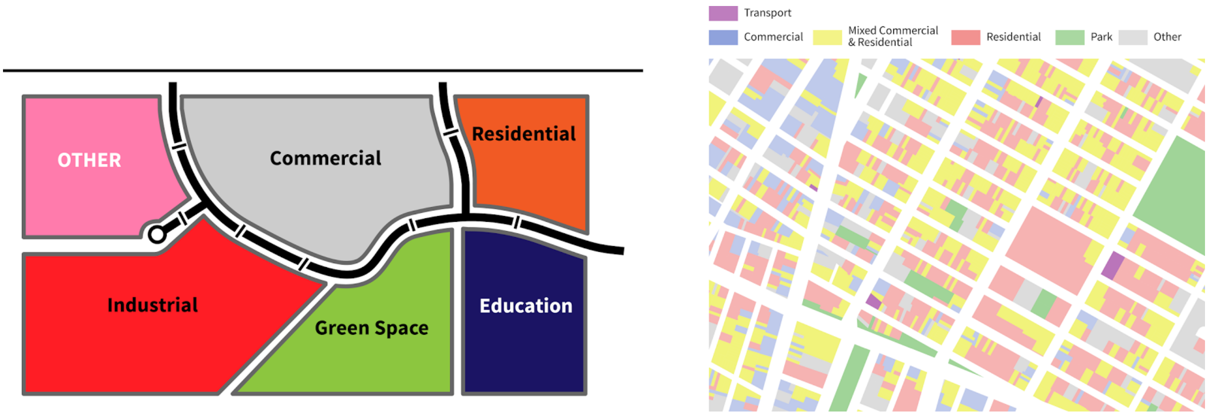


Figure 6 - Single use zoning vs multi use mixed zoning.

However, this kind of planning which promoted single use zones proved to be poor at creating walkable communities and disastrous for providing efficient transport (Winkelman et al. 2010) and promoted the car to the primary transport mode. Jane Jacobs is most famous for her attack on this planning style and its segregationist principals. Her writings were driven by an objection to the regeneration plans being laid out by New York’s then Transport Commissioner, Robert Moses (Flint. 2009).

Jacob's argued for the neighbourhoods of New York City, where she observed a socially diverse and mixed land use. It was these which she argued (Jacobs, 1961) are what shape vibrant communities. In *Life and Death of Great American Cities* (Jacobs, 1961) she detailed the key characteristics of vibrant neighbourhoods. The main themes which influence this are, Economic & Social Vitality, the power of physical design, which included the provision of narrow crowded multiuse streets, higher densities of people, the removal of single use zoning, and the redesign of streets for people not cars (Wickersham, 2001).

More recently urban theory has shifted towards the sustainable city, with a renewed focus on 'the walkable city' (Southworth, 2005). The same factors which can create vibrant neighbourhoods also work towards the more sustainable city. Transportation is a key part of achieving this. *"Every journey in a city begins and ends with a walk"* (Speck, 2014) - this also holds true for any trip using the bike system.

New York City is ranked as the most walkable city in America according to Walkscore (2014). The city of New York presents an ideal opportunity to compare the results of relationship strength to identify communities within a city. It contains many diverse neighbourhoods comprising various mixes of land uses, varying densities, urban design, economy and people. Primarily it should be noted that *"what is barely hinted at in other American cities is condensed and enlarged in New York."* (Bellow, 1970/1994) This along with the availability of large and open datasets were key considerations in choosing New York City for this study.

4. Data Sources

The New York City bike share system operator CitiBike publishes a number of datasets on their website. The data which will be used in the study will comprise of the monthly origin-destination trip data from July 2013 - August 2014 (CitiBike, 2014) relating to the ridership, station location and capacity in addition to the station locations feed.

4.1. Station Data

There are 332 bike share stations in New York City. These are located mainly on the island of Manhattan and the borough of Brooklyn. There are over 6000 bikes which make up the system. The location of each bike station can be seen in the following image.

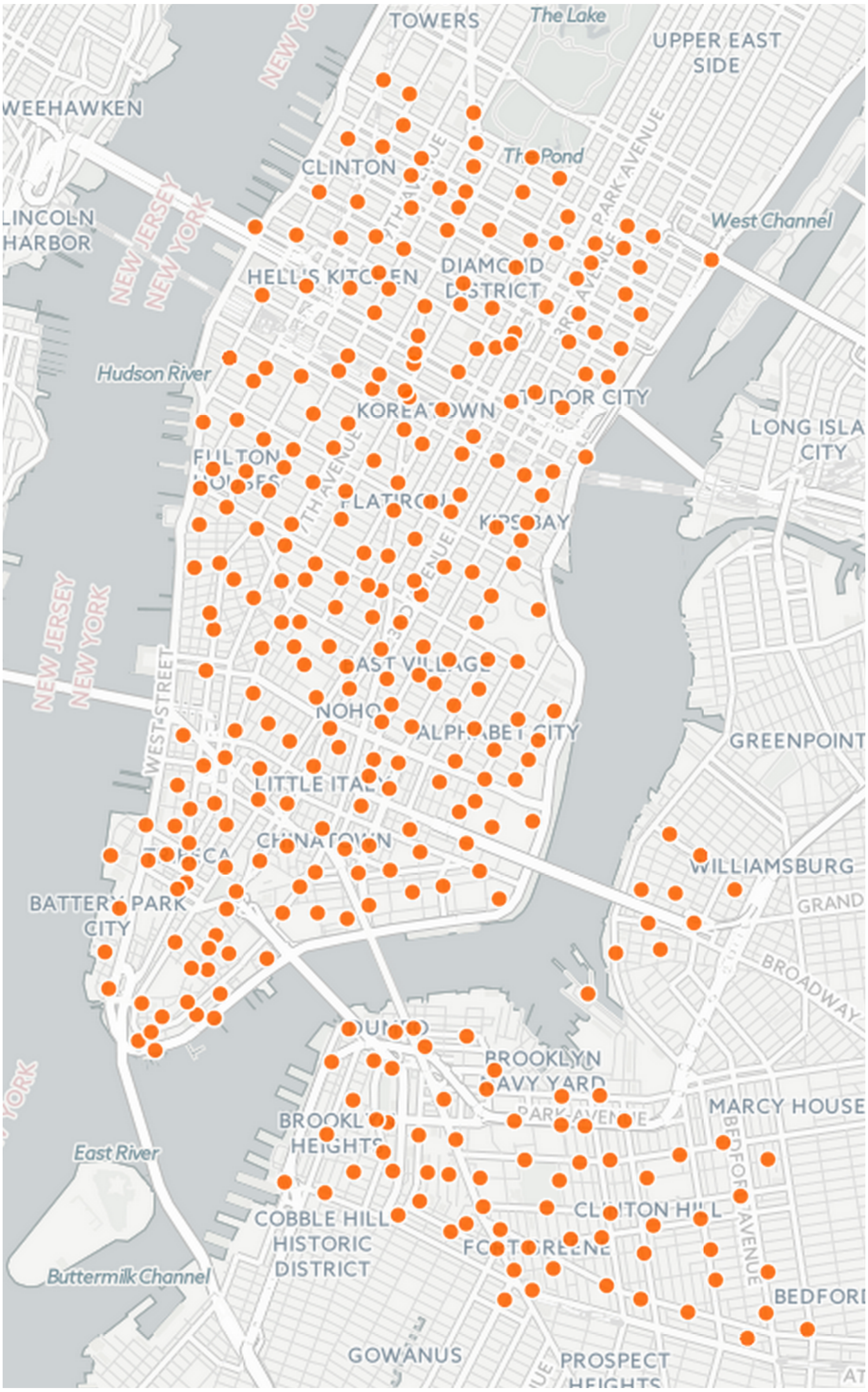


Figure 7 - NYC Bike Station Locations (CitiBike, 2015)

The station data feed is published as a JSON (ECMA, 2013) (JavaScript Object Notation) feed. It contains the the following fields for each station, along with sample values:

```
"id": 72,
"stationName": "W 52 St & 11 Ave",
"availableDocks": 34,
"totalDocks": 38,
"latitude": 40.76727216,
"longitude": -73.99392888,
"statusValue": "In Service",
"statusKey": 1,
"availableBikes": 3,
"stAddress1": "W 52 St & 11 Ave",
"stAddress2": "",
"city": "",
"postalCode": "",
"location": "",
"altitude": "",
"testStation": false,
"lastCommunicationTime": null,
"landMark": ""
```

This is a live feed and as such one can see the number of bikes which are available at the station at this moment. For the purposes of this study only the `id`, `station name`, `latitude` and `longitude` fields were used. Each station has a unique `id` number.

The station data feed can be accessed at - http://www.citibikenyc.com/stations/json. (CitiBike, 2015)

4.2. CitiBike Origin-Destination Trip Data

CitiBike also publishes data relating to all of trips made by people using the bikes since the system opened in May 2013. It publishes this data under an open license which is fairly uncommon. This data is the main data source for this dissertation.

This data is published in CSV (Comma Separated Value) format at the end of each month. Each file contains all the trips made within the period of one month. For the purposes of this study trips from July 2013 - August 2014 have been used. This amounted to some 10, 407, 546 (CitiBike, 2014) individual bike trips during this period.

The structure of the data fields is as follows:

- `Trip Duration (seconds)`
- `Start Time and Date`
- `Stop Time and Date`
- `Start Station Name`
- `End Station Name`
- `Start / End Station ID`
- `Start / End Station Lat/Long`
- `Bike ID`
- `User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member)`
- `Gender (Zero=unknown; 1=male; 2=female)`
- `Year of Birth`

The main fields of interest for use in this study are `start time`, `end time`, and the `start and end station ids`.

This origin-destination data can be accessed at - http://www.citibikenyc.com/system-data. (CitiBike, 2014)

The table below shows some summary statistics relating to the origin-destination trip data.

Number of Trips	10,407,546	
Time Period	July 2013 - 31 August 2014	
Max Distance Interstation (Quickest Route)	14.66 km	
Longest Trip	72 Days (151) Cleveland Pl & Spring St (501) FDR Drive & E 35 St	Start: 2013-07-08 16:51:40 End: 2013-09-19 01:10:53
Shortest Trip	1,471 trips of 1 minute or less.	

Table 1 - Summary Statistics for NYC CitiBike bike share scheme

4.3. Neighborhoods of New York

The New York City Department of City Planning each year publishes a map detailing the different neighbourhoods of the city. The city is divided into *"5 boroughs, 59 community districts and hundreds of neighborhoods"*. (Department of City Planning, 2014) These boundaries are developed by combining census parcels to form neighbourhood tabulation areas. An extract of the map can be seen below with the bike stations layered on top for context.

Figure 8 - Map of Neighbourhoods covering bike share scheme area.

This data is delivered in a shapefile format and contains the following fields:

BoroCode - Borough Reference Number

BoroName - Borough Name

CountyFIPS - County Reference Number

NTACode - Neighborhood Tabulation Area Code

NTAName - Neighborhood Tabulation Area Name

Shape_Leng - Perimeter in feet

Shape_Area - Area in square feet

This shapefile can be downloaded at the following link - http://www.nyc.gov/html/dcp/html/bytes/dwn_nynta.shtml (Department of City Planning, 2014)

4.4. OpenStreetMap Data - New York City

It was deemed appropriate to use street data from OpenStreetMap to perform the routing queries given its wide spread coverage for the chosen city of New York. This data was acquired through the "Metro Extracts" tool by Mapzen which outputs weekly extracts of many cities throughout the world. (Mapzen, 2015)

The data is provided in a number of different formats. The format used in this research is the "OSM" format. This provides the data in a zipped file containing an **.osm** planet file. This is an XML file which *"contains a list of instances of [OSM] data primitives (nodes, ways, and relations) that are the architecture of the OSM model."* (OpenStreetMap Wiki, 2015)

The daily updated planet file can be accessed here -
https://s3.amazonaws.com/metro-extracts.mapzen.com/new-york_new-york.osm.bz2
(Mapzen, 2015)

4.5. NYC Dept. Planning PLUTO Database

Given the evidence by Saelens et al. (2003) and the overall contribution of mixed land uses to creating higher desirability and community in different parts of the city, it is considered extremely important to consider the land uses types in the vicinity of the bikeshare stations. To do this the Property Land Use Tax Lot Output (PLUTO) dataset from NYC Department of City Planning is used.

The PLUTO dataset contains geographical features which are "*derived from the Tax Lot Polygon feature class which is part of the Department of Finance's Digital Tax Map (DTM)*." (NYC Department of City Planning, 2014) The dataset features a number of attributes derived from the planning departments PLUTO database which include, information on current land use, number of floors, current zoning designation, etc. The current land use for each parcel is of primary interest as it will then be possible to identify if the area within the immediate vicinity of a bike share station is dominated by a single land use or is highly mixed.

The data is provided in a number of shapefiles. The entire database is disaggregated to the boroughs of New York for ease of distribution. The files used by this research are those which cover both Brooklyn and Manhattan.

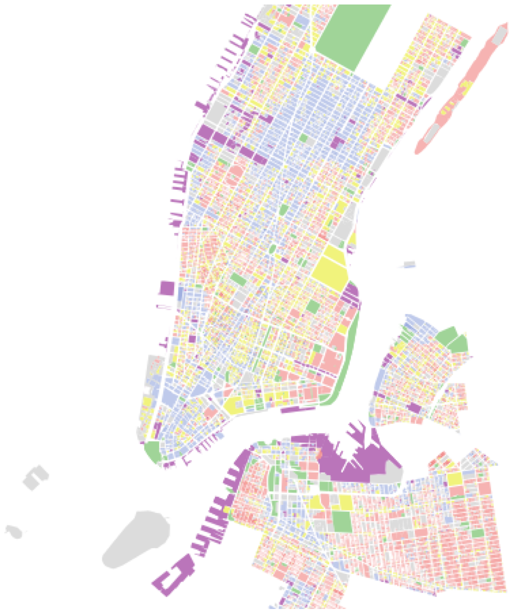


Figure 9 - Extract of data from NYC MapPLUTO dataset with colours categorising the landuse of that parcel.

5. Methods

As mentioned previously, all data used in this study is available open source and with open licenses. In the spirit of *"open research"* or *"open science"* (Woelfle, 2011) all of the code developed to complete this study is open and freely available for reuse through Github.

Given this approach to open research, it was felt appropriate to write the methods section of this dissertation in a similar style to a user manual. This will allow any researcher to simply follow the steps as set out below to reproduce or build upon this research.

As such this section is divided into two distinct parts: the first part gives a brief overview of the entire procedure to be followed; and the second part provides full specific details on each steps which are to be followed when reproducing this research.

5.1. Overview of complete procedure

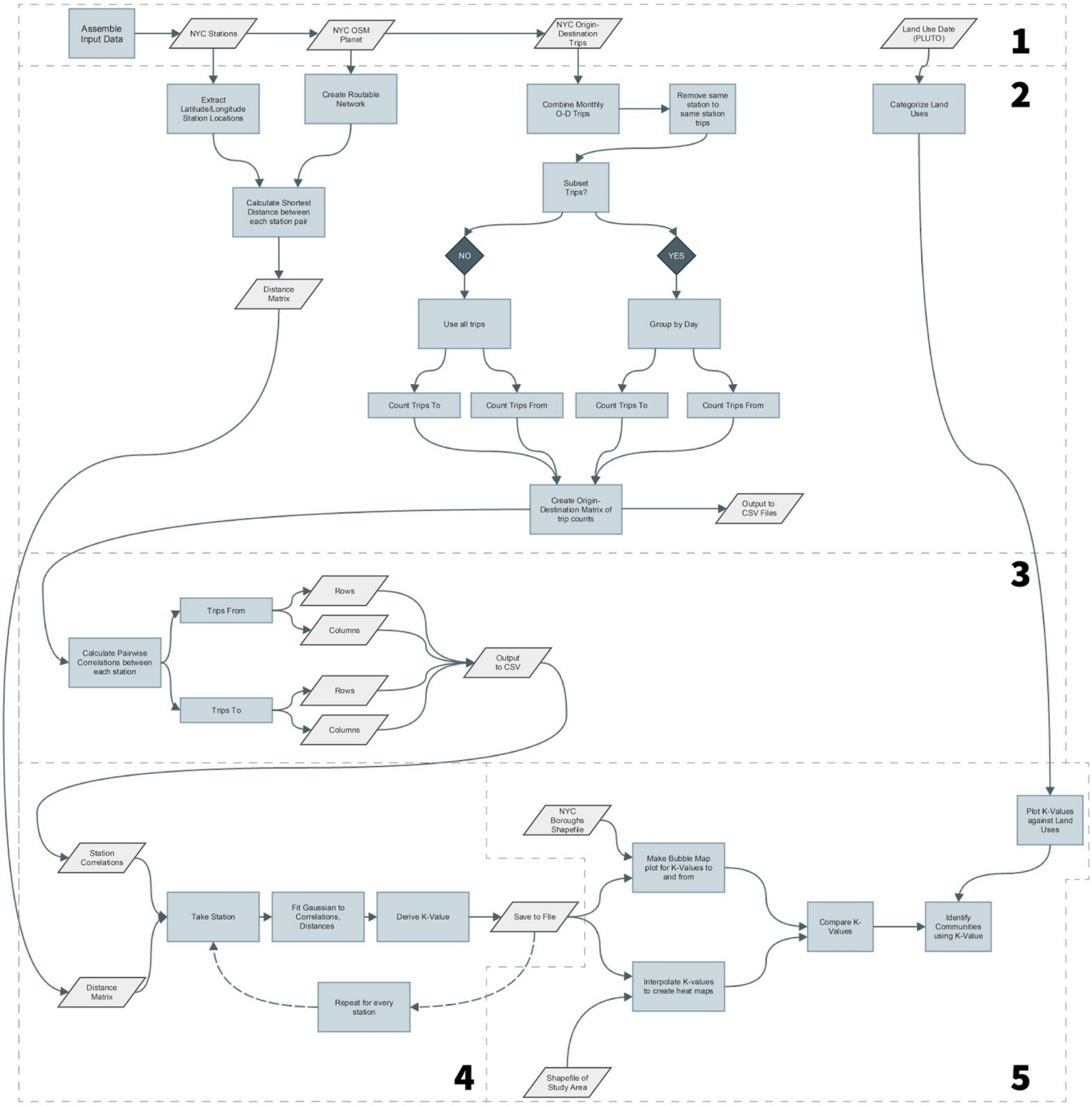


Figure 10 - Overview of processes used to extract K-Values. (Large format version located at Appendix 2)

The overall procedure for conducting this research can be viewed in process diagram form at Figure 10 above. A large format version of the procedure is provided at Appendix 2 for clearer readability. The procedure can be broken down into five key parts.

The first part involves gathering all of the input data sources to conduct the research. In the second part the shortest bicycling distances are extracted between each station, trip count origin-destination matrices are made, and New York City land uses are categorised.

The third part extracts the correlations for each station. The correlations and distance matrix are then combined, the distance decay function fitted, and then K-Values are derived for each station. Finally in the fifth part these K-Values are mapped, compared against land uses in their vicinity, and analysed.

5.2. Gather the Data

The first step is to gather all of the required data by downloading the following:

- Clone the github repository containing the code used to perform the analysis.
- The repository contains the bike stations file including longitude and latitude coordinates.
- Origin-Destination data for required period of time from Citibike.
- Obtain a copy of the most recent OSM Planet File for New York.
- Download shapefile of NYC from repository.

5.3. Extract the data

Each of the origin-destination trips files `YYYYMM-citibike-tripdata.zip` must be extracted to a folder of your choice. The only requirement for running the later scripts is that these trip files are all contained within the same folder. For example to `../data/citibike`

5.4. Get Interstation Distances

Given that the main aim is to extract intrastation pairwise linear correlations for each station pair and relate this to distance, it is imperative to get the distance between each station pair.

For the sake of speed in reproducing this research these distances have been provided in the research repository in the file `station_dists_nyc.txt`. This is a CSV file containing the `start_station_id`, `end_station_id` and `distance`.

To complete this step you require two data sources, which are as follows:

1. `./data/new-york_new-york.osm.bz2`
2. `./data/station_latlons_nyc.txt (id,lat,long,name)`

The `station_latlons_nyc.txt` file is already located in the data folder of the repository.

The scripts used by this step are as follows:

1. `./src/getStDists.py`
2. `./src/router.py`

We only need to run `./src/getStDists.py` as this calls `router.py` to get the distance.

```
> getStDists city = nyc <city=london/nyc>
```

The script will then check to see if the OSM planet file has been split. This checks that the XML OSM planet file has been converted into a format which can be used for routing.

`Router.py` first extracts the .osm file from zip file. It then proceeds to prepare the file for use with Routino for finding the shortest path between each station pair. *"Routino is an application for finding a route between two points using the dataset of topographical information collected by http://www.OpenStreetMap.org."* (Bishop, 2014)

In OpenStreetMap, roads and paths are stored as **`ways`**, which are referenced with the tag **`higway=*`**. Using the `planetsplitter` function of Routino, we can create a network graph which will be used by Routino for getting the shortest path between each station by traversing the street network (edges).

A configuration file is used to tell `planetsplitter` which highway tags should be extracted from the osm file and what weight to apply to that edge containing that way on the network graph.

On completion the following files are generated:

- `ny-nodes.mem`
- `ny-relations.mem`
- `ny-segments.mem`
- `ny-ways.mem`

The next function call is `router.getAllNodes(nyc)`. This function reads the `.osm` planet file and extracts all of the nodes contained within it. This is used later as an input for the routing function.

The next step is to get the geographical bounds of the `.osm` file. As part of the schema of the planet file, this is included in a section tagged **bounds**. This contains two tuples which reference the min and max values of the bounding box as follows: `(minlat, minlon)` `and (maxlat, maxlon)`.

```
> getLatLons (city="nyc")
```

This will read in the latitude and longitude values of the bike stations from `../data/station_latlons_nyc.txt`. and return these as an array containing the fields `id, lat, lon`.

A check is then done to make sure that the station latitude and longitude values are within the bounds of the `.osm` file. If this is the case, the function `.writeDMat()` is called to calculate the distances for each station pair.

```
> writeDMat (latlons, nodes, city)
```

The results are written to `../results/station_dists_nyc.txt`

This will take each pair of station latitude & longitude coordinates and find the “quickest” route between them. This makes use of the router function of Routino. The router function can be customised by options set in two configuration files..

routino-profiles.xml

This file contains the configuration relating to how Routino will weight each edge of the graph. This contains a number of different profiles for different modes of transport such as foot, bicycle, wheelchair, motorcar, hgv, etc. Each profile specifies maximum speeds and types of highway which can be traversed using the specified mode.

`routino-translations.xml`

This file is used by Routino to provide language translations for turn directions, highway tag description, etc. It is also used to configure which output files are generated by routing.

The router is passed the following switches:

--transport = bicycle

--quickest (find the quickest route between two latlong coordinates)

--lon1 = first longitude coordinate --lat1 = first latitude coordinate

--lon2 = second longitude coordinate --lat2 = second latitude coordinate

The router will take roughly 1 - 2 seconds to return a route between the two points. It will then generate five files containing information relating to the *"quickest"* route. The files are as follows:

- `quickest.html`
- `quickest-track.gpx`
- `quickest-route.gpx`
- `quickest.txt`
- `quickest-all.txt`

Once the router is called it will generate a HTML file called `quickest.html`. This file contains a map, a list of directions waypoint by waypoint from the first station to the second station, and the total distance of the route. A web scraper is then used to read the distance from this file.

An example of this can be seen in the image below.

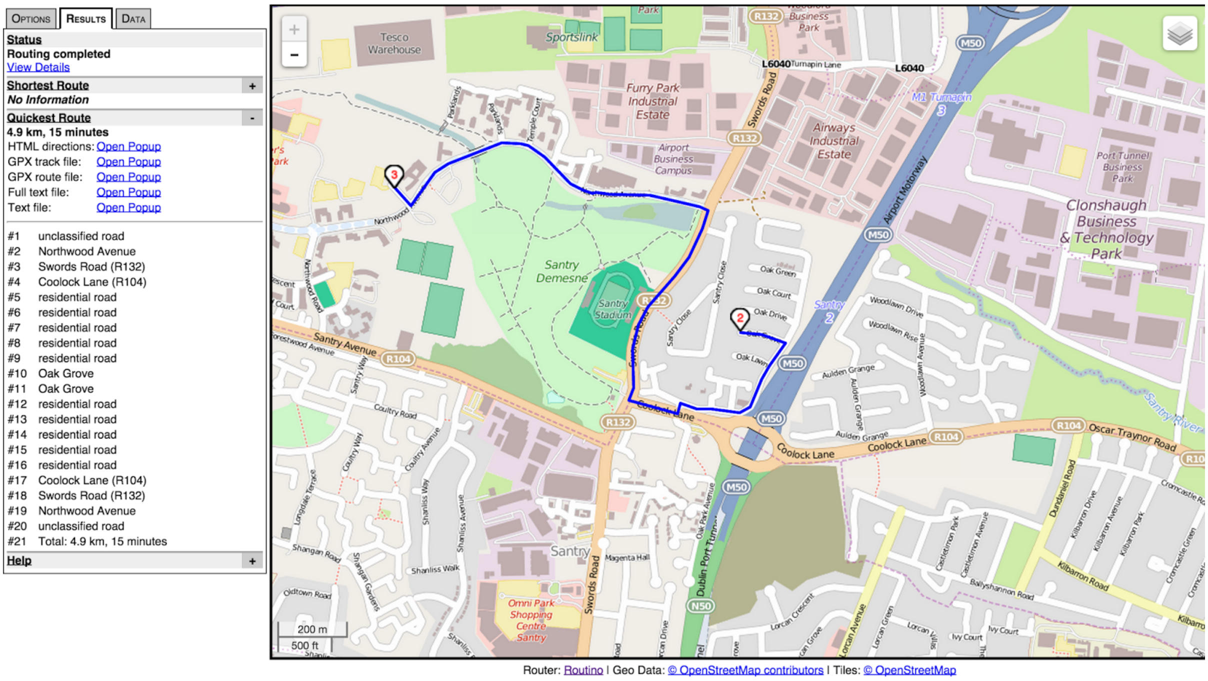


Figure 11 - Routino quickest.html routing file output

The start_station_id, end_station_id, and distance are then written to the output file `../results/station_dists_nyc.txt.`

This process is then repeated for each station pair so that a distance is calculated from each station to every other station. For efficiency the script will find the distance between each pair only once in one direction.

5.5. Calculate the number of trips between stations

The number of trips between each station will be used as the value for which to calculate the pairwise correlations between each and every station, so a method to extract the counts from the trip files is provided in the script `get_trip_counts.py`.

```
> python ./python/get_trip_counts.py -city -directory-of-trips-files -stations-file -period
```

This script makes use of the features provided by the `pandas` (Python Data Analysis Library) library for Python.

The data files required to run this script are as follows:

1. `../data/nyc_usage_stats/` (location of extracted origin-destination trips files)
2. `../data/nyc_station_latlons.txt`

When running this script a number of options can be provided:

--city specifies which city to use for the counts. currently nyc or london
--period specifies the time period on which to perform the counts. currently total & weekday

All of the trip files contained in the specified directory will be read in one at a time and checked for consistency. The *start_id, end_id, start_time,* and *end_time* will be read from each of the trip files.

Once they have been read in, they will then all be combined into a single `pandas.dataframe`. This single data frame is then indexed by the `start_time`. This allows us to use pandas inbuilt functions for querying data frames by a time index which is useful for extracting the daily counts.

Pandas allows the use of SQL (Standard Query Language) style queries to subset data. There are a number of requirements when counting the trips. These are as follows:

1. To exclude trips which start and end at the same station;
2. To group the results by the start and end stations;
3. To return counts for the number of trips beginning (from) each station;
4. And, to return counts for the number of trips ending at each station (to).

The `df.query()` function is used to make SQL like queries. Next the `df.groupby()` function is used to group the results by a certain column. This returns the data frame with a multi array index. Finally of the `group.size()` function is used, which will aggregate the results of the grouping by counting the number of rows in each. One row is one trip. To do this use the query below:

```
counts_from_each = df.query('start_id != end_id and end_id != start_id').groupby(['start_id',"end_id"]).size()
```

For instance as an example one can take the first 5 stations, and the results will be returned as follows:

start_id	end_id	count
79	72	182
79	82	25
79	83	4
79	116	140
82	72	16
82	79	35
82	83	7
82	116	6
83	72	20
83	79	35
83	82	12
83	116	90
116	72	105
116	79	221
116	82	5
116	83	43

Table 2 - Result of running count query for trips from the first five stations.

Next this flat data frame is transformed into a square matrix by using the `.unstack()` function. The square matrix will have the `station_ids` as both columns names and row names, like an origin destination matrix. It should be noted that there are no values for trips from the same station to the same station.

start_id	72	79	82	83	116
72					
79	182				
82	16	35			
83	20	35	12		
116	105	221	5	43	

Table 3 - Result of unstacking count for trips from first five stations.

Once unstacked, one is only interested in the lower triangular of the counts from and the counts to. We use the Python Numpy (Numpy developers, 2014) library to extract the `lower.triangular()` of each data frame and copy this to the upper triangular so that the matrices returned are square. The columns and rows are then indexed by `station_ids`.

The same process is used to get the counts of specific subsets of the data. As indicated above, we can subset the counts by weekday. We do this by adding an extra column to the data frame indicating the day of the week the trip took place, ranging from 0 to 6.

Pandas has inbuilt functions for indexing time series. As mentioned above we indexed the data frame containing all the trips by start time of each trip. This allows the use of the `index.weekday()` function from pandas to insert an integer into the new interval column of our data frame. This indicates the day of the week using the following index from 0 - 6:

```
0 = Monday
1 = Tuesday
2 = Wednesday
3 = Thursday
4 = Friday
5 = Saturday
6 = Sunday
```

The query for the counts is then modified to use an additional parameter to query the interval column. A loop is used to get the counts for each day of the week, starting with Monday.

```
counts_from_each = df.query('interval >= %s and interval <= %s and
      start_id != end_id and
      end_id != start_id' % 0).groupby(['start_id',"end_id"]).size()
```

In both cases two data frames are returned from the counting function, one containing counts of trips from, and the other counts of trips to each station. These are then written to two CSV files:

```
../results/nyc_total_from.csv
../results/nyc_total_to.csv
```

These two files will then be used for calculating the pairwise correlations.

5.6. Pairwise Correlations

The next step is to calculate the pairwise correlations for each station using the results from the previous step which returned trip counts.

```
> python ./python/reg1.py -city nyc
```

This script will read in the trip counts and calculate the pairwise linear correlations.

The data files required to run this script are as follows:

1. `../results/nyc_total_from.csv` **or** `nyc_`**`weekday`**`_total_from.csv (mon, tue, wed, etc.)`
2. `../results/nyc_total_to.csv` **or** `nyc_`**`weekday`**`_total_to.csv (mon, tue, wed, etc.)`

The **`-city`** argument used to call the script is simply used to indicate which prefix to use when reading in the CSV files. `nyc_trips_from.csv` and the `nyc_trips_to.csv` files are first read into pandas as separate data frames.

This script makes use of functions from the python library SciPy, which is *"a collection of open source software for scientific computing in Python"*. (SciPy developers, 2013) We make use of the linear regression functions provided by `scipy.stats`. This function *"computes a least-squares regression for two sets of measurements"*. (SciPy community, 2014)

The correlation is calculated by running `scipy.stats.linregress()` using a row and the next row. This is repeated with the same row and each other row. The same process is then repeated for each and every other row. A correlation coefficient `(r_value)` is returned, which is then squared. This is repeated for every row in the matrix. This is also completed against columns.

This results in two square matrices, one for the row correlations, and the other for the columns. These matrices have their columns and rows indexed by the station ids. The coefficient calculation is performed for both the trips_from and trips_to.

The results are saved to the following files:

- `../results/nyc_total_from_cols.csv`
- `../results/nyc_total_from_rows.csv`
- `../results/nyc_total_to_cols.csv`
- `../results/nyc_total_to_rows.csv`

This script can also be run against the trip counts resulting from using the period argument.

All correlations computed are positive between $0 < r < 1$.

5.7. Make Distance Matrix

In order to perform the analysis in the next step, which is to get the maximum distance of the correlation. First the `nyc_station_dists.txt` file is transformed into a square matrix. At the moment this file is in the form of a list, an extract can be seen below:

```
“start_id”,”end_id”, “dist”
72,72,
72,79,6.43
72,82,7.458
72,83,11.546
72,116,3.784
```

```
72,119,10.252
72,120,12.78
```

Table 4 - Extract from station distances file

```
> python make_dist_matrix.py
```

Using python pandas library, the script reads in the list and sets a multi index using the `start_id` and `end_id`.

Then again using `df.unstack()` the data frame is transformed/pivoted into a square matrix.

This is then saved to a CSV file: `../data/nyc_d_matrix.csv`

5.8. Combine coefficients with distances

For the final steps in the analysis we will move from python to R. For the next stage we need to combine the coefficients with the distance between each station. It will also fit a function to the distribution of each stations coefficient in order to predict at what distance that correlation will have dissipated.

```
> R -f ./python/nyc1.R
```

This script will first read in each of the correlation CSV files which were outputted in the last step.

- `../results/nyc_total_from_cols.csv`
- `../results/nyc_total_from_rows.csv`
- `../results/nyc_total_to_cols.csv`
- `../results/nyc_total_to_rows.csv`

We then read in the station distance file as an R `data.frame`, `nyc_station_dists.csv`. Next a check is made to make sure that the lower triangular of the matrix is the same as the upper triangular. This should be a square distance matrix, so both triangulars should match.

The Nonlinear Least Squares (`nls`) function from the R stats library will be used to fit a function to the correlations and distances.

The distance matrix and correlation matrix are first transformed to the vectors **`ndm, rvec`**. Both vectors are then compared to remove any missing values. The distance vector is then transformed using a logarithmic scale of base 10.

`k` is defined as the value of the distance at which the maximum distance for which the station is correlated to any other station. This is not a gravity based model as is traditionally used by the field of economics but rather a distance decay model.

The nls model formula is defined as follows with two constants, cc and a. It is expected that when `cc = 0`, so that `rvec -> 0` as `d` also becomes very large as a result. In all cases `cc < 1`.

```
rvec ~ cc + a * exp(-((d - min(d))/k)^2)
rvec = correlations as vector
d = distance vector
cc = constant
a = constant
k = K-value
```

It should be noted that the above formula is somewhat modified because the main aim of the research is to discern differences between areas/community hubs, to a "log-Gaussian" distance decay model because it yields more pronounced spatial variation.

Since we know that the `from_columns, from_rows, to_columns` and `to_rows` files are pairwise correlations, the lower triangular and the upper triangular of the matrices are the same. We can run the nls model against either the rows or the columns of each file and expect the same results for either to or from file.

Given the sensitivity of `nls` to starting parameters, the R function `TryCatch{}` is used to allow for different values to be used as starting parameters.

```
get_mod <- function(rvec, d, as, kss, tr) {
      tryCatch(nls (rvec ~ cc + a * exp(-((d - min(d)/k)^2)),
      start=list(cc=0, a=as,k=kss), trace=tr), error=function(e) NULL)
 }
```

If the model returns `NULL` , this means it will stop if the model does not converge or produces an error. It will then try the next set of start parameters. The different values used by the model are detailed below.

cc	a	k
0	1	2
0	0.2	2
0	1.76	0.38
0	0.8	0.5
0	0.8	0.6
0	0.8	0.3
0	0.187	2.505
0	0.8	1.2
0	2	1
0	1	mean(d)
0	0.2	1

Table 4 - Starting Parameters used for `nls.mod`

The resulting **`K-values`** are merged into a data.frame containing the station names and latitude longitude coordinates. Once merged these are then saved to CSV files as follows:

- ../results/kvals/nyc_tc_k.csv
- ../results/kvals/nyc_tr_k.csv
- ../results/kvals/nyc_fc_k.csv
- ../results/kvals/nyc_fr_k.csv

5.9. Map K-value Results

This final step is to designed to make visual plots of the results of the various analyses performed above. There are two types of visualizations which can be produced: bubble plots and kriging interpolations.

5.9.1. Bubble Plots

```
> R get_bubbles.R
```

Bubble maps showing the magnitude of K-values at the stations throughout the city are created using the functions contained in the `get_bubbles.R` script. This script makes use of the ggplot2 graphics package for R.

The files required for this script are as follows:

- ../data/study_area/study_area_neighbourhood_boundaries.shp
- ../results/kvals/nyc_mon_tc_k.csv
- ../results/kvals/nyc_mon_fc_k.csv
- ../results/kvals/nyc_tue_fc_k.csv
- ...etc.

The K-values are read in from the CSV files for both to and from trips for the period specified. For instance the Monday K-values to and Monday K-values from are read in as **`mon`** and **`monf`** variables.

The boundaries file containing the neighbourhoods in the study area are read into a `spatial.data.frame` and assigned a map projection. Using the `fortify()` function from ggplot2 (Wickham, 2009), the `spatial.data.frame` is coerced into a `data.frame` of ordered points so that ggplot can be used to plot the graphics.

The plots are outputted with both the to and from values on the same graphic. As such both to and from variables must be passed to the `make_plot` function as follows:

```
> make_plot(mon, monf, '1_Monday', kvals)
```

This function takes the input data.frame containing the stations file with latitude, longitude coordinates and K-values. It creates two `ggplot()` graphics in which the size and colour of the points representing K-values are determined by binned K-values.

```
k_size <- scale_size_continuous(
            name="K-Value (km)", range=c(1,10), limits=c(0,25),
            breaks = c(0,3, 5, 7, 9, 11,13, 20, 25),
            labels=c(0,"< 3", 5, 7, 9, 11,13, 20, "25 +")
            )

k_colour <- scale_color_gradient(
            name="K-Value (km)", high="darkorchid3",low = "orange",
            limits=c(0,25),
            breaks = c(0,3, 5, 7, 9, 11,13, 20, 25),
            labels=c(0,"< 3", 5, 7, 9, 11,13, 20, "25 +")
            )
```

A standard legend and scale are used for the graphics to allow for results to be compared between the overall results and those for each day of the week. **`Limits`**, is used to mask the outlier values of k that result over the weekends. **`Breaks`**, bins the values of *K-values* into 9 ranges, and labels simply gives names to these ranges for use in the plotting of the legend. **`Range`**, is the sequence of numbers that ggplot uses to set the size of each point depending on the range of the K-values set by the breaks.

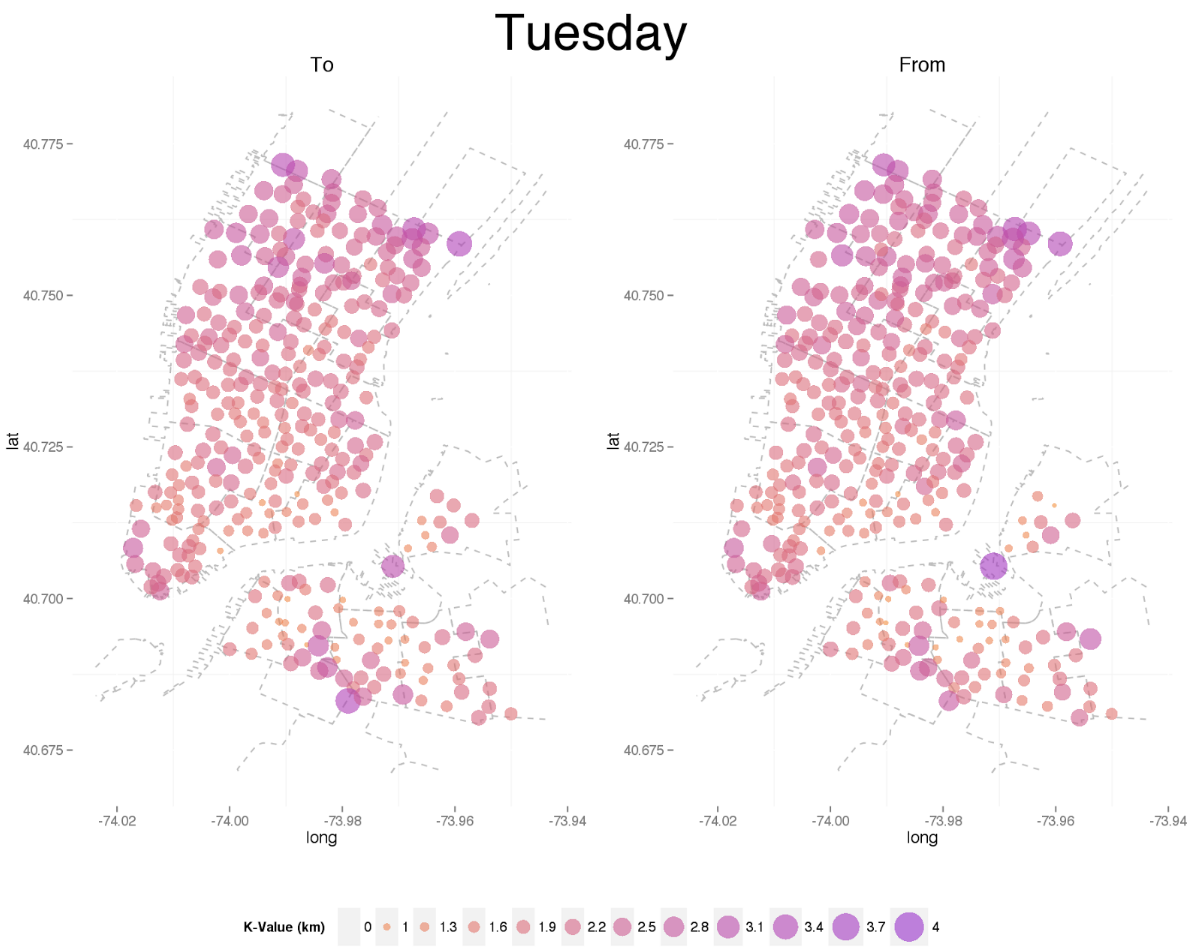


Figure 12 - Example bubble plot for all trips made in New York City on Tuesdays.

5.9.2. Interpolation / Heat Map

The other method of mapping the K-value results was to use a method to estimate the values of k across the study area.

An outline shape for the island of manhattan and the brooklyn were extracted from openstreetmap data. This shape is then used to mask the kriging area.

This step uses R's spatial (sp) package once again. The stations data.frame is then converted to a SpatialPointsDataFrame. The study area polygon is also converted to a SpatialPolygonDataFrame.

Coordinate transformations were performed on both the station points and the study area shape to transform coordinate reference system from WGS84 (EPSG:4326) to NAD83 /

UTM zone 18N (EPSG:26918). The main reason for this is so that we can set up our interpolation grid in units of meters rather than degrees.

A grid upon which to interpolate estimated values to is created with 50m x 50m cells. This grid is then clipped to the shape of Manhattan and the study area using polygons from OSM data. `shape/nyc1.shp` is included in the repository. This was chosen as the smallest distance between two stations is **131m**.

The actual kriging is done using the `autoKrige()` function from the R package automap (ordinary kriging). (Hiemstra et al., 2008) The formula supplied to this function is as follows:

```
> autoKrige(log10(kval)~1, ks_proj, clip_grid)
```

`ks_proj`: SpatialPointsDataframe of station K-values
`clip_grid`: the polygon grid to interpolate to

The predicted values are then plotted using the `spplot()` function from the sp package. The resulting interpolated maps are similar to the following, **Figure 13**. The main aim of this step is to help the visual interpretation of the resulting K-values. This will help in identifying the hotspots of high K-values.

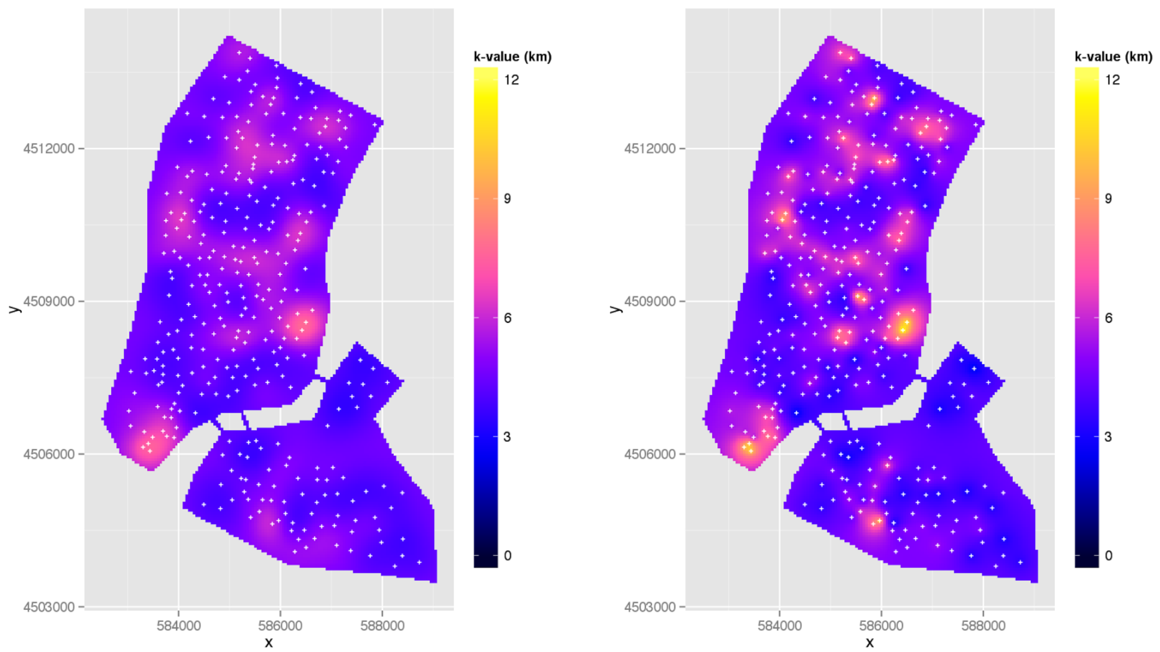


Figure 13 - Interpolated Map of New York K-values

5.10. Map K-values and Land Uses

For this step a suitable GIS such as QGIS should be used. The two shapefiles containing the PLUTO data should be loaded first into a suitable database such as a geo-enabled PostgreSQL database using the POSTGIS extension. This will allow the easy generalization of land uses classifications. The commands for loading this data can be found in `commands_to_run.md` file of the repository.

The field of interest from the PLUTO dataset is called `landuse`. This field contains an integer ranging from 1 - 11. The number references the primary land use on each tax parcel area.

<table>
<tr><th>Code</th><th>Land Use</th><th>Generalised To</th></tr>
<tr><td>01</td><td>One and Two Family Buildings</td><td rowspan="3">Residential</td></tr>
<tr><td>02</td><td>Multi-Family Walkup</td></tr>
<tr><td>03</td><td>Multi-Family with Elevator</td></tr>
<tr><td>04</td><td>Mixed Residential & Commercial</td><td>Mixed Residential & Commercial</td></tr>
<tr><td>05</td><td>Commercial & Office</td><td rowspan="2">Commercial</td></tr>
<tr><td>06</td><td>Industrial & Manufacturing</td></tr>
<tr><td>07</td><td>Transport & utility</td><td>Transport</td></tr>
<tr><td>08</td><td>Public Facilities & Institutions</td><td>NA</td></tr>
<tr><td>09</td><td>Open Space & Recreation</td><td>Open Space & Recreation</td></tr>
<tr><td>10</td><td>Parking Facilities</td><td>NA</td></tr>
<tr><td>11</td><td>Vacant Land</td><td>NA</td></tr>
</table>

Table 5 - PLUTO Land Use classes

These land uses can then be mapped in the GIS. The resulting K-value CSV files can then be overlaid on these maps. Distinct colours should then be assigned to each land use class. The most important classes being Residential, Mixed Residential & Commercial, and Commercial.

6. Results & Discussion

The methods described in the previous section were applied to the New York City bike share data, which is subsetted by a number of time periods. The complete analysis was run for each day of the week as well as the complete dataset.

It is considered appropriate to describe and discuss the results for each time period separately in order to highlight the resulting patterns from each time period or subset. Finally the results will be analysed together to identify any patterns which are represented through the different subsets.

The resulting maps will then be overlaid with the Neighbourhoods of New York City boundaries map to determine whether any of the communities present from the cycling patterns are aligned with those as identified by the city. Finally, the same process will be completed by overlaying the K-values for trips over the PLUTO land use map.

It should be noted that from here forward when using the word community or hotspot, refers to an area in which there are a number of stations with the similar K-values.

6.1. Distance Decay Model Fits

As detailed in the Theory section four different distance decay models were tested against the resulting correlations and distances. Each of these models is compared by their results SSR values for two sample stations. The resulting model fit for each of these is shown at Figure 14 alongside the SSR values.

Station	S1	S2
Gaussian	1.78	1.71
Exponential	1.87	1.92
Inverse	4.69	3.95
Gravity	5.86	4.07

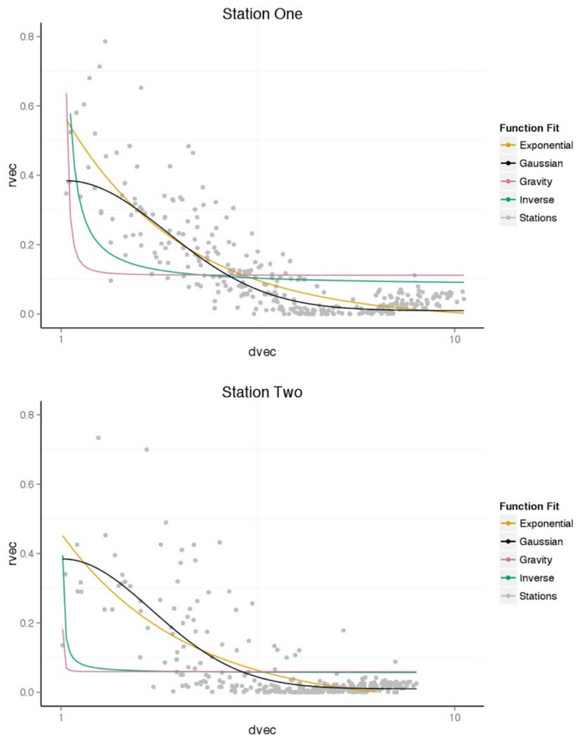


Figure 14 - Comparison of fitted distance decay models on two sample stations and resulting sum of squared residuals.

As can be seen in the table Figure 14 above, the Gaussian function is the one which shows the lowest sum of squared residuals, closely followed by the exponential decay model. For both stations the inverse and gravity models perform poorly in terms of their respective fits to the real data. These results further demonstrate that a Gaussian decay model is more suited for short distance trips as found by Padgham (2012) than those of traditional economic decay or power law models.

Furthermore a comparison was made to ascertain whether a "log-Gaussian" distance decay model, the model formula implemented in the methods section, would further help in discerning the differences between areas. K-values were derived using both direct- and log-distances for both trips to and trips from. The comparison used to test the spatial variation was SD/mean. Log-distances showed a higher variation (0.393) meanwhile direct-distances showed a lower variation (0.201).

6.2. Complete Trip Dataset

To begin with, the entire trip's dataset was run through each step of the method first. Below are some key statistics regarding the data which was analysed:

Period of Data	July 2013 - August 2014
Total Number of Trips	10, 134, 537 *(excl. same station/same station)*
Trips From	10, 134, 537
Trips To	10, 134, 537
No. of Bike Stations	332

Table 6 - Complete Trip Dataset Key Statistics

This first run of the data produced a number of interesting results. Principal among these findings is that, non intuitively, none of what one would consider as the key destinations to visit in New York emerged as hotspots, such as major transport nodes or business districts as was the case in previous studies (Austwick et al. 2013, Padgham, 2012).

The resulting K-values for each station are plotted at Figure 15, below. As can be seen, the two areas which are identified with the largest filled circles, representing the K-values (strength of relationship to stations in vicinity), are located on the South Western part of Manhattan adjacent to a large social housing area, and to the West where the Williamsburg Bridge joins to Manhattan.

For every step the trips are split between trips to each station from every other station, and trips from each station to every other station. You can clearly see that the pattern of the trips doesn't really change depending on whether the trips are either "from" or "to". A paired t-test was performed and although the two datasets (trips from, trips to) were significantly different (p-value: 0.00651) mean differences in k-values were in fact very small (mean: 2.027) corresponding to only 1% of mean values. This is further evidence that the k-values for trips from do differ significantly from trips to as was observed by Padgham (2012) in relation to fixed spatial pattern in London.

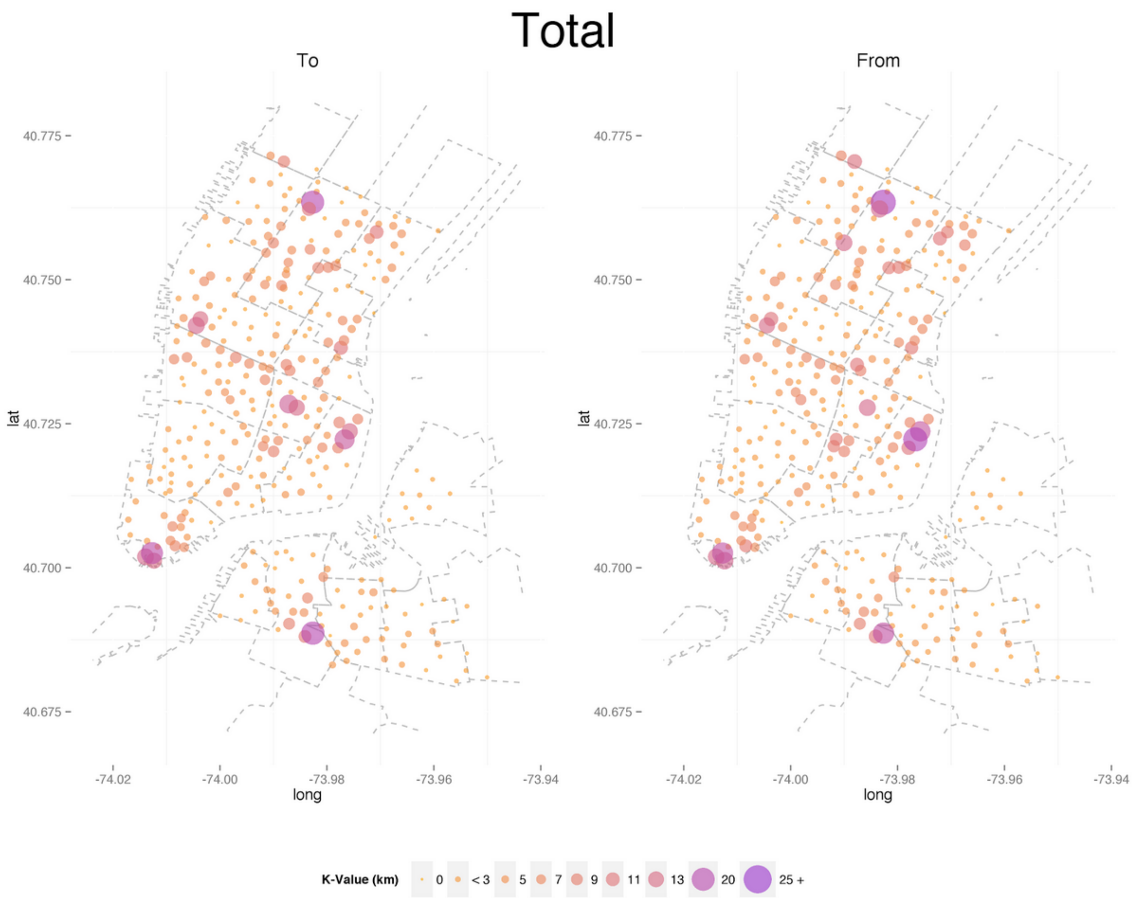


Figure 15 - NYC Trips From and Trips To K-Values

Because they are statistically independent, the two datasets (to- and from-) provide an opportunity to compare two different estimates of k-values. When compared visually, it is clear that the same stations which have high K-values with regard to trips from, also have high values of trips to.

The areas surrounding these stations can be thought of as key community hubs within the study area as significant volumes of travel originate or terminate at these nodes. Furthermore, the effective distance of activity generated by these stations extends much further than those surrounding them. It is expected that these areas will contain a mix of land uses (Frank, L.D. and Pivo G., 1994).

The stations which have higher values are activity centres for movement in both directions, to and from that station. Interestingly when overlaid with the New York City boroughs, almost every single one shows at least one group of stations with high K-values. In both cases there are at least five separate communities.

Community hubs can clearly be identified, characterised by the large purple circles, in the following boroughs: nearby well known destinations are identified in brackets.

- Lower East Side (Jacob Riis Houses / Social Housing)
- Battery Park City-Lower Manhattan (South Ferry Subway / Ferry Terminal)
- Hudson Yards-Chelsea-Flatiron-Union Square (Fulton Houses / High Line)
- Midtown-Midtown South (broadway/west 53rd, west 51st)
- East Village
- Gramercy
- DUMBO-Vinegar Hill-Downtown Brooklyn-Boerum Hill (Nearby Long Island University / Department of Work)
- Lincoln Square (Fordham University Lincoln Campus / Amsterdam Houses)

When further investigated through the review of city maps showing the area surrounding the stations which show very high K-values, it can be seen that there are a number of key municipal services nearby. These act as trip generators for travel between community hubs (Hanson, S., Schwab, M., 1987). For instance, the New York Department of Labour is less than 200m from the community in South Brooklyn.

Station Area Name	**PLUTO Land Use Character**	**Key Services**
Hanover Pl & Livingston St	Commercial and Mixed Residential & Commercial	NYC Department of Labour, Long Island University, NYC Department of Housing, Library.
South St & Whitehall St, Bus Slip & State St and Water - Whitehall Plaza	Mixed Residential & Commercial, Park and Commercial.	South Ferry Terminal
E 6 St & Avenue D	Residential, Mixed	Jacob Riis Housing Project,

	Residential & Commercial and Park	Church, New York City Housing Office
Avenue D & E 8 St	Residential, Mixed Residential & Commercial and Park	Jacob Riis Housing Project, Church, New York City Housing Office
St Marks Pl & 1 Ave	Mixed Residential & Commercial, Residential	Cafes, Bookshops, nearby social housing project (200m).
9 Ave & W 16 St	Commercial and Mixed Residential & Commercial	Chelsea Markets, Google, High Line, New York City Housing Project
Broadway & W 53 St	Residential, Commercial, Park and Mixed Residential & Commercial.	Broadway Theatres

Table 7 - PLUTO land use character adjacent to stations with highest k-values.

These activity centres are located next to universities such as Long Island University in Brooklyn, New York City Department of Housing social housing projects such as Jacob Riis Houses (Lower East Side) and Fulton Houses (Hudson Yards), and key tourist attraction areas such as St Marks Place (East Village), the theatres of Broadway between 51st and 54th Streets (Midtown) and the South Ferry Terminal (Battery Park).

In a number of cases these key stations seem to located in close proximity to key government services or government sponsored amenities for communities; the NYC Department of Labour offices in Brooklyn; the food markets at Chelsea Market adjacent to the Google Offices, and parks such as The High Line.

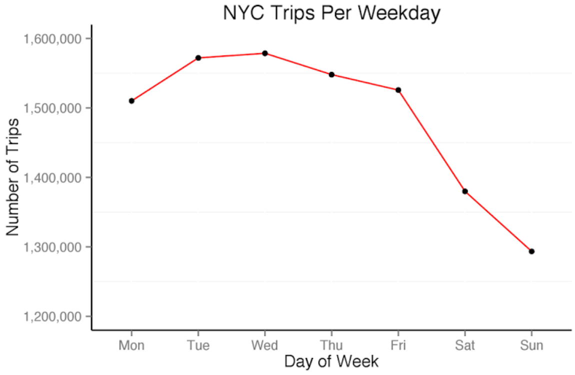
What is clearly absent from these key areas, not as one would expect, is the lack of major transport hubs being identified in the results. One would assume that such key transport nodes such as Penn Station and the key business areas surrounding World Trade Centre would be highlighted. In previous studies comparing different cities with bike share systems key transport hubs anchored movements, see Austwick et al. (2013) and Lathia et al (2012) . It should be noted that each of these locations identified have bike stations in their vicinity.

From the initial analysis, it is clear that only the areas which have a high number of people living there feature strongly when compared to those which are primarily work orientated.

6.3. Weekdays

Previous studies have all identified that there is generally a seasonally dependent variation in usage pattern on the bike share systems. (Austwick et al., 2013. Padgham, 2012. Lathia et al., 2012.) This is particularly evident when viewing the differences between workdays (Monday - Friday) and weekend days (Saturday - Sunday). Given this variation it was deemed appropriate to examine the patterns of K-values over each day of the week. The day for which each trip took place was identified and the dataset was then partitioned by the day of the week.

It can be seen at Figure 16, that the total number of trips each day increases until Wednesday and then falls away to its lowest level on Sundays. We might expect then in the resulting K-value plots that different centres should emerge over the weekend when compared with the weekdays. There are approximately 290,000 less trips made on Sunday compared to Wednesday.



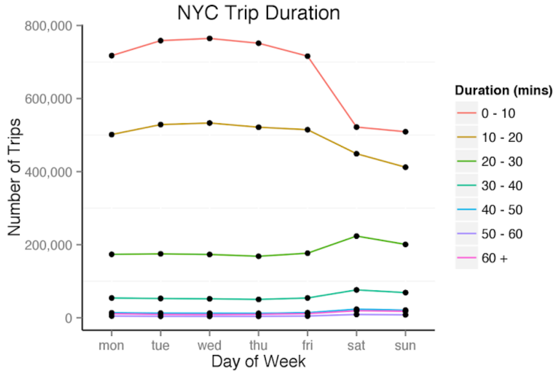


Figure 16 - Total Trips per weekday. Number of trips grouped by journey trip duration per day

Similarly, while the number of trips per day decreases on the weekend days of Saturday and Sunday, the duration of those trips increases. Given this behaviour one could assume that trips at the weekend, while taking longer, may also travel further than those during the week. The maximum shortest route distance between any of the station pairs is 14.67 km. However, one would assume this journey could be much longer.

If one examines the top 10% of K-values throughout the week one can see that the maximum values increase up to Wednesday, the day of highest trip volume (Figure 16). There is a second spike on Saturday when the highest number of rides of 30-40 minutes duration occur. Sunday also appears to be similar to Saturday.

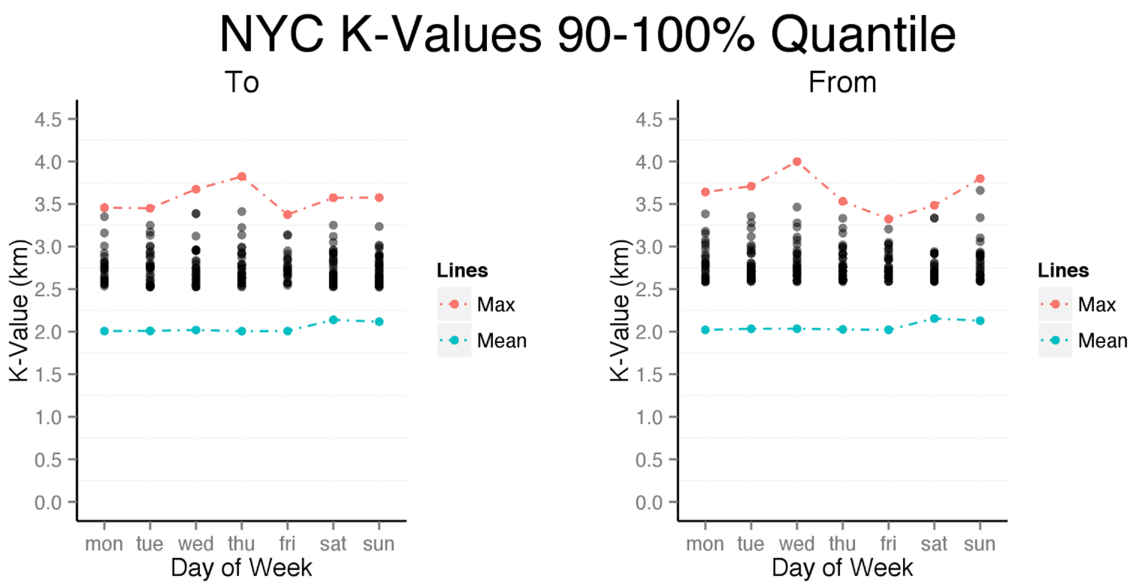


Figure 17 - NYC K-Values for the 90% and 100% quantiles.

All the stations were plotted on a map where the K-value is represented by the diameter of the point. This was done for each day of the week. An animation was prepared in order to examine the changes in K-values visually. This can be viewed at Figure 17.[2] The trend of the K-values increases until Wednesday, then falls off before it peaks on Saturday as can be clearly seen.

The same communities which were identified in the results produced by the total trips can also be identified here. Visually examining the changing points, Saturday and Sunday definitely present a different dynamic when compared to the weekdays which further reflects the temporal variations found in travel patterns between weekdays and weekends of Austwick et al (2013) and Padgham (2012) research.

[2] See http://cbroderick.me/thesis/

The tourist destinations of East Village, South Island Ferry Terminal, Broadway, New York City Library all show with higher K-value at the weekend. This could indeed be a sign that at the weekend the community pattern leans more towards culture and leisure activities (Austwick et al. 2013).

The same stations that present high K-values on Monday continue to do so throughout the week. Some show a minor reduction over the weekend, however this could simply be attributed to the greater number of higher K-values on Saturday and Sunday.

Figure 18 - New York K-Value dynamic throughout the week, Monday to Saturday.

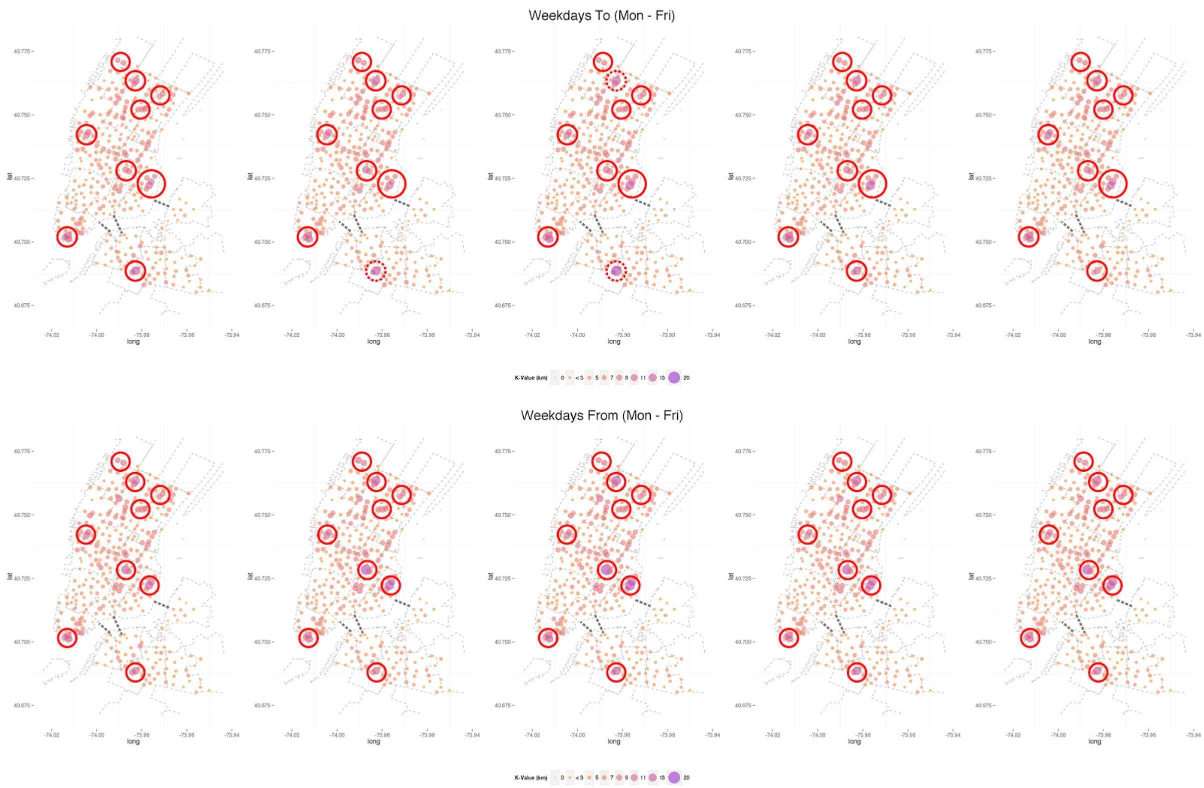


Figure 19 - New York K-Values on Weekdays (Monday - Friday) with high K-value communities identified

The community that is identified in Brooklyn nearby Long Island University does not show with such high K-values on Tuesday and Wednesday when looking at trips to. Meanwhile, these same communities show for each weekday when looking at trips from only.

During the week there are clearly up to nine centers which have groups of high K-values. Contrast this to the weekend, Figure 20, where there are up to ten different communities. This suggests that there is a higher variety of key destinations and focal points of activity than during the week. As mentioned previously, when the area surrounding these locations identified are examined, it is found that the majority of these are located in proximity, less than 500m or five minutes walk, to key cultural and tourist centres.

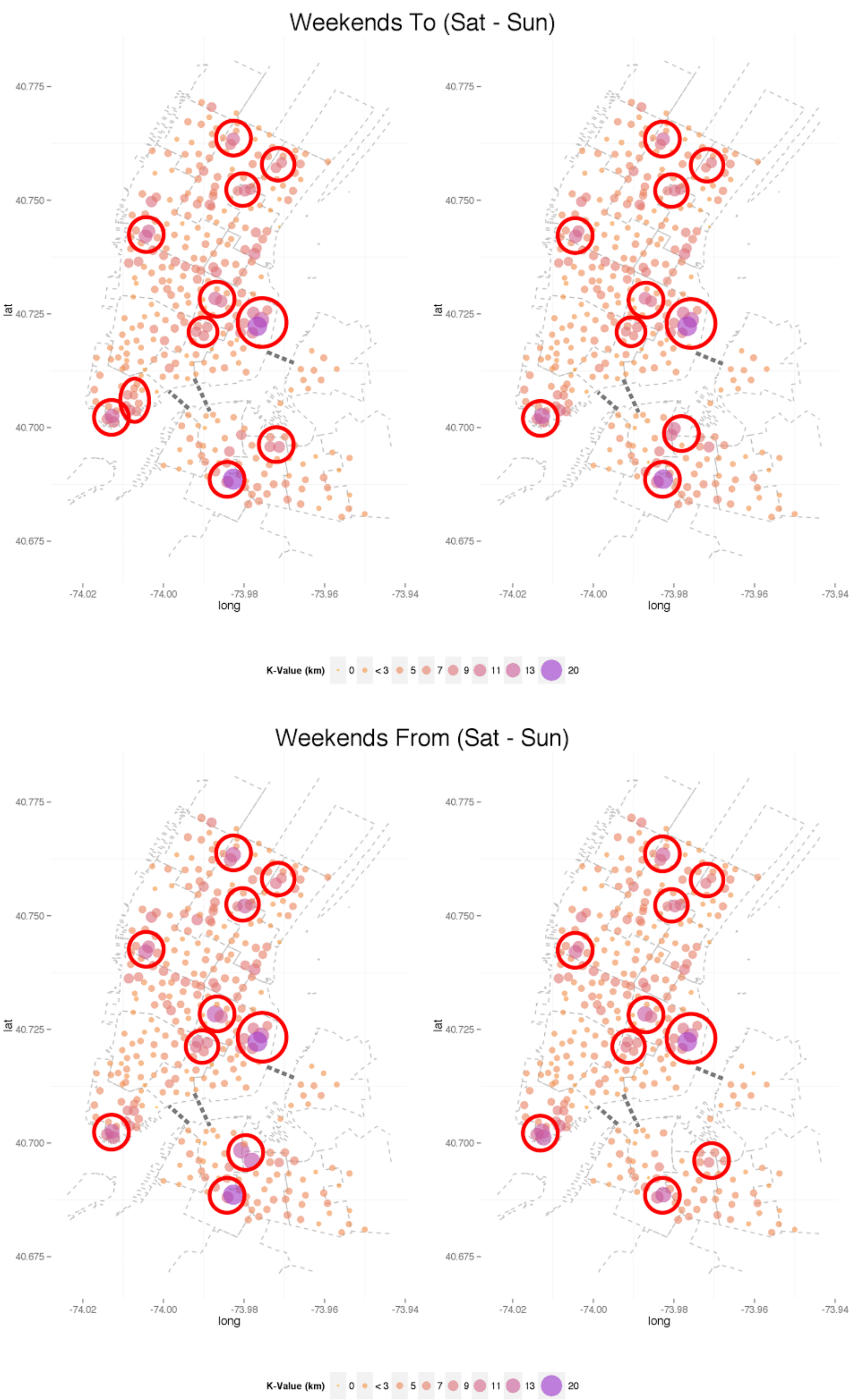


Figure 20 - New York K-Values on Weekends (Saturday - Sunday) with high K-value communities identified

Following this, the key centres were then mapped to determine their proximity to certain land uses within the city. For this the PLUTO database was used, as the current land use of each tax parcel is identified. In this case the main land uses of interest were residential, commercial and mixed use. PLUTO provides 13 land uses and for the purpose of this study these were generalized, as follows:

Residential: One and Two Family Buildings, Multi-Family Walkup and Multi-Family with Elevator.
Commercial: Commercial & Office, and Industrial & Manufacturing
Mixed: Mixed Residential & Commercial

All communities identified in Figure 20, which have high K-values are located in areas where the predominant use according to PLUTO is either Residential or Mixed Residential & Commercial. The exception to this being those stations located adjacent to South Ferry Terminal and those nearby the New York Public Library and Grand Central Station, where the use is mostly commercial.

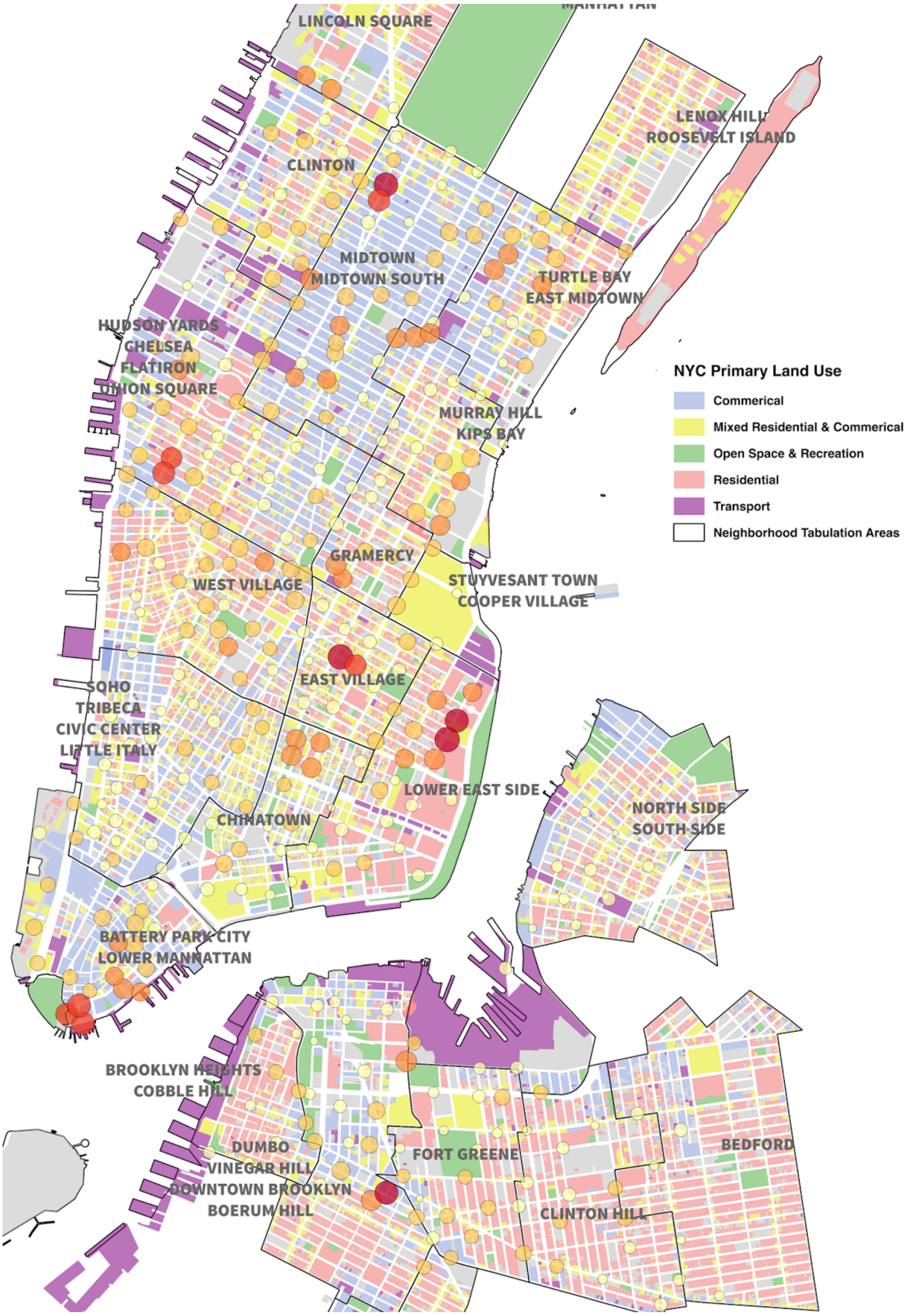


Figure 21 - New York total trips from K-Values plotted against PLUTO land use types.

Table 7, details the primary land uses found in the vicinity of the top nine stations ranked by magnitude of their respective K-values. The areas which have clusters of high K-values are also the areas where there are a mix of land use in close proximity to the bike stations.

These are the real places where people live and work. These are the same neighbourhoods where there are numerous social and cultural destinations. People need to move form from place to another to utilise different types of services (Berger and Meijers, 2012) as highlighted by the range of government services found in the different community hotspots.

What is most striking is that the communities which are within the commercial areas (blue) above, are located near their edges, where they blend back into the mixed residential commercial communities. Walking and cycling will be higher in these areas due to the fact there is no single dominant land use as shown by Frank, L.D. and Pivo G. (1994) and Saelens et al. (2003) especially when compared to the other main transport modes.

Each of the community hubs in essence acts as a monocentric centre where at its heart is a group of key services, such as governmental services, which serve to attract movement from surrounding hubs. These centres act independently as observed by Berger and Meijers (2012) promoting higher or lower desirability as reflected in the difference in K-values present at each station.

Padgham (2012) found that approximately 80% of movement within cities follows a similar spatial pattern at any time. This can also be seen here for New York City when viewing the areas of high k-values throughout the week, which do not drastically differ throughout the week, see Figure 17. A pairwise comparison (21 in total) was made between each day of the week and it was found that on average 72% of the total k-value variance is explained by a fixed spatial pattern in New York City.

This work is important for further understanding the concentrations of movements within cities. The urban development models of spatial planners have been taken for granted and not rigorously tested. This work builds upon previous work to understand movement within the city on a micro level (Austwick et al. 2013. Padgham 2012 and 2014.).

The use of a Gaussian distance decay model is unique in it's application by this study, in that unlike gravity based models they have well defined means and variances, and can thus be interpreted in direct spatial terms. This allows the measure of the similarity of movement for every point within a transport system. It is very important to know where these hotspots of both high and low k-values are as they can provide indicators of the vitality and overall levels of activity within certain areas of our city. Ultimately these K-values can serve as one of the first steps to being able to empirically define what the boundaries of a neighbourhood or community is in the context of the complex spatial network that is the city.

7. Future Research

There is potential for future research in this area to focus on replicating this study in other cities which have extensive bike share systems such as London or Paris. Similarly this type of analysis could be combined with other transport modes such as the metro or bus to rule out any possible bias of the bike data.

The relationship between the K-value for each station and the types of community, land uses within their vicinity along with the cultural importance should be further explored. During the course of this research there was not sufficient time to statistically test this owing to the difficulty in grouping adjacent land uses. This should be tested as part of future research into this area.

To expand the use of this kind of analysis on other data sources which contain large volumes of short distance trips such as smart card trip data, car share trip data or taxi data. In the case of smart card data, for example in London there are over 3.5 million trips per day (Transport for London, 2014) and this will require the extra initial step of calculating the origin destination pairs due to its use on different modes of transport. Once this is complete the trips can then be analysed using the methods contained in this research.

8. Conclusions

In conclusion this research shows that it is indeed possible to derive a measure to describe the cultural significance of a city area to that city using small scale travel data. These k-values demonstrate that the strength of the relationship between the stations of the New York City CitiBike scheme and their nearby stations are related to the land uses that surround them.

In pursuit of delivering both open science and reproducible research, this this thesis is an example of one way in which this can be achieved using a combination of open source tools and software with open data.

All previous studies of this kind from available literature have measured a single/global K-value for the whole city. The focal point of this research has been to derive this strength of the relationship between stations for each point within the city.

Furthermore, this research demonstrates that using these K-values in combination with current land use data shows that this measure of importance corresponds with areas of where there is no single dominant land use type.

This supports the idea that valued locations within large cities are ones that are planned for variety and diversity.

Limitations

This study has focused on the New York City where the transport system is more complex than simply the bikeshare system located in the central part of the city. It does not consider the other modes of transport at play including subway, ferry, bus or taxi.

The study are similarly is confined to the island of Manhattan and parts of Williamsburg and Brooklyn due to the location of the bike system being only in these areas. As such the patterns of travel observed may differ if it were possible to increase the size of the study

The methods developed for this dissertation were only applied to New York City on data from the twelve month period of July 2013 to August 2014.

The primary means of identifying the predominant land uses within the vicinity of bike stations, this could be further improved by developing a model to accurately quantify the dominant uses in the vicinity.

Finally this study does not seek to identify community hubs based on the socioeconomic makeup of these areas but merely based on their respective travel patterns.

Bibliography / References

Anas, A. Arnott, R. Small, K.A. (1998). “Urban Spatial Structure”. Journal of Economic Literature. 36, pp. 1426-1464.

Atlanta Regional Commission. (2010) "Mixed-Use Development - the Atlanta Regional Commission." Available at: http://www.atlantaregional.com/File%20Library/Local%20Gov%20Services/gs_cct_mixedusetool_1109.pdf [Last Accessed: 16 Jan. 2015]

Australian Bureau of Statistics. (2013) “Statistical Language - Correlation and Causation”. Available at: http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+correlation+and+causation. [Last Accessed 03 February 2015].

Banister, D. (2008) “The sustainable mobility paradigm”. *Transport Policy*. Volume 15, Issue 2, March 2008, Pages 73–80. doi: dx.doi.org/10.1016/j.tranpol.2007.10.005

Batty, M. (2013) The New Science of Cities. Cambridge, MA: *MIT Press*.

Bellow, S. (1970/1994) `New York: World-Famous Impossibility ', New York Review of Books. Reprinted in It All Adds Up. London: *Secker & Warburg*.

Bishop, A.M. (2014). “Routino : Router for OpenStreetMap Data.”Available at: http://www.routino.org/. [Last Accessed: 2 January 2015].

Bolker, B.M., and BT Grenfell. (1996) "Impact of vaccination on the spatial correlation and persistence of measles dynamics." *Proceedings of the National Academy of Sciences* 93.22 (1996): 12648-12653.

Brantz, D. and Dümpelmann, S. (2011). Greening the City: Urban Landscapes in the Twentieth Century. Charlottesville: *University of Virginia Press*. Pg. 17.

Burger M., Meijers E. (2012). “Form follows function? Linking morphological and functional polycentricity”. *Urban Studies* 49(5): 1127–1149.

Cheng, I-H., and Howard J.WI. (1999) "Controlling for heterogeneity in gravity models of trade and integration." *Federal Reserve Bank of St. Louis Working Paper Series* 1999-010 (1999).

CitiBike. (2014) System Data. Available at: http://www.citibikenyc.com/system-data. [Accessed 01 December 14].

CitiBike, (2015) Station Map. Available at: http://www.citibikenyc.com/stations [Accessed 01 January 15].

CitiBikeNYC (2014) How it Works. http://www.citibikenyc.com/how-it-works. [Last Accessed: 27, July 2014]

CitiBikeNYC (2014) System Data. http://www.citibikenyc.com/system-data. [Last Accessed: 27, July 2014]

City Planning Department City of New York (2014) Pluto and MapPluto. http://www.nyc.gov/html/dcp/html/bytes/dwn_pluto_mappluto.shtml. Last Accessed: 27, July 2014

CitiBike, (2015), Station Feed JSON. Available at: http://www.citibikenyc.com/stations/json [Accessed 01 January 15].

CitiBike. 2014. System Data. Available at: http://www.citibikenyc.com/system-data. [Accessed 01 December 14].

Correlation *Oxford English Dictionary.* (2012) 7th ed. Oxford: Oxford University Press.

Department of City Planning, New York City. (2014) "City of Neighborhoods - New York City ... - NYC.gov." Available at: http://www.nyc.gov/html/dcp/html/neighbor/neigh.shtml [Last Accessed: 2 Jan. 2015]

Department of City Planning, New York City. (2014) *"Neighborhood Tabulation Areas - BYTES of the BIG APPLE - New York City Department of City Planning - NYC.gov."* Available at: http://www.nyc.gov/html/dcp/html/bytes/dwn_nynta.shtml [Last Accessed: 2 Jan. 2015]

ECMA International. (2013) “The JSON Data Interchange Format.“ Geneva: *Standard ECMA-404*.

Flint, A. (2009). “Wrestling with Moses: How Jane Jacobs Took on New York's Master Builder and Transformed the American City”. New York: *Random House*.

Frank L.D., Pivo, G. (1994) “Impacts of mixed use and density on utilization of three modes of travel: Single-occupant vehicle, transit, and walking.” *Transportation Research Record*. 1994, 1466:44–52.

Jacobs, J. (1961) “The Death and Life of Great American Cities.” New York: *Random House*.

Hanson S, Schwab M. (1987). “Accessibility and intraurban travel.” *Environment and Planning A.* 1987, 19:735–748.

Hiemstra, P.H., Pebesma, E.J., Twenhofel, C.J.W. and G.B.M. Heuvelink. (2008) Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network. *Computers & Geosciences*, accepted for publication.

“Introducing JSON” (2003) Online. Available at: http://json.org. [Last Accessed: 2 January 2014]

Kloosterman, R. Musterd, S. (2001) “The Polycentric Urban Region: Towards a Research Agenda”. *Urban Studies*. 38(4), pp. 623-633.

Lathia N, Ahmed S, Capra L (2012) Measuring the impact of opening the London shared bicycle scheme to casual users. Transportation Research Part C: Emerging Technologies 22: 88–102. doi:10.1016/j.trc.2011.12.004.

Lockwood, D. (2007) "Mapping crime in Savannah social disadvantage, land use, and violent crimes reported to the police." *Social Science Computer Review* 25.2: 194-209.

Meddin, R., DeMaio, P. (2014) The Bike-Sharing World Map. Available at: http://bit.ly/K9pKmO. [Last Accessed: 19/12/2014]

Midgley, P. (2009) “The role of smart bike-sharing systems in urban mobility”. *JOURNEYS.* 23–31.

Mapzen. (2015) “Mapzen - Start where you are.. 2015.” Available at: https://mapzen.com/metro-extracts/. [Last Accessed: 1 January 2015].

Mokhtarian, P., Salomon, I., Handy, S. (2006) “The impacts of ICT on leisure activities and travel: a conceptual exploration”. *Transportation*, 33 (3) (2006), pp. 263–289

Numpy developers. (2013). “NumPy — Numpy.” Available at: http://www.numpy.org/. [Last Accessed: 2 January 2015].

NYC Department of City Planning. (2014) “New York City, MapPLUTO 14V2.” Available at: http://www.nyc.gov/html/dcp/pdf/bytes/meta_mappluto.pdf. [Last Accessed: 12 January 2015]

Odland, J. (1988) Spatial autocorrelation. Newbury Park, CA: Sage Publications. P. 7

Onnela, Jukka-Pekka et al. (2011) "Geographic constraints on social network groups." PLoS one 6.4: e16939.

OpenStreetMap Wiki. (2015) “OSM XML - OpenStreetMap Wiki.” Available at: http://wiki.openstreetmap.org/wiki/OSM_XML. [Last Accessed: 2 January 2015].

Padgham, M (2014) Sorted partitions and urban structure. Journal. Unpublished.

Padgham, M. (2012) "Human movement is both diffusive and directed." *PloS one* 7.5: e37754

Park, R., Ernest W. B. and Roderick D. McK. (1925) *The City*. Chicago: University of Chicago Press.

Saelens, B., Sallis, J.F., Frank. L.D. (2003) “Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures”. *Annals of Behavioral Medicine*. (2003) 25(2):80–91. doi: http://doi.org/10.1207/S15324796ABM2502_03

Schnare, A.B., and Raymond J.S. (1976) "Segmentation in urban housing markets." *Journal of Urban Economics* 3.2: 146-166.

SciPy developers. (2013). “Scientific Computing Tools for Python.” Available at: http://www.scipy.org/about.html#scientific-computing-tools-for-python. [Last Accessed: 2 January 2015].

Smyth, G.K. (2002) "Nonlinear regression." *Encyclopedia of environmetrics* 4. Pg. 1406. doi:10.1002/9780470057339.van017.

Speck. J. (2014). “The general theory of walkability | Jeff Speck | TEDxMidAtlantic.” [Online Video]. 23 May. Available from: https://www.youtube.com/watch?v=uEkgM9P2C5U. [Last Accessed: 04 January 2015].

The Scipy community. (2014). “scipy.stats.linregress — SciPy v0.14.0 Reference Guide.” Available at: http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.linregress.html. [Last Accessed: 2 January 2014].

Southworth, M. (2005). ”Designing the Walkable City.” *J. Urban Plann. Dev. 131, SPECIAL ISSUE TITLE: Innovating Regulations in Urban Planning and Development*. 246–257.

Tobler, W.R. (1970) "A computer movie simulating urban growth in the Detroit region." *Economic geography.* 46(2). Pg: 234-240.

Taylor, N., (1998) Urban planning theory since 1945, *SAGE Publications Ltd*, London. doi: http://dx.doi.org/10.4135/9781446218648.

F.J. van Lenthe, , J. Brug, J.P. Mackenbach. (2004) “Neighbourhood inequalities in physical inactivity: the role of neighbourhood attractiveness, proximity to local facilities and safety in the Netherlands”. *Social Science & Medicine*. Volume 60, Issue 4, February 2005, Pages 763–775 doi: http://dx.doi.org/10.1016/j.socscimed.2004.06.013

Weisstein, E.W. (2014) "Least Squares Fitting." MathWorld--A Wolfram Web Resource. Available at: http://mathworld.wolfram.com/LeastSquaresFitting.html [Last Accessed: 18 Janruary 2014]

Walk Score. (2015). “Most Walkable Cities in the United States, Canada, and Australia on Walk Score.” Available at: https://www.walkscore.com/cities-and-neighborhoods/. [Last Accessed: 19 January 2015].

Wickham, H. (2009) ggplot2: elegant graphics for data analysis. *Springer* New York.

Wickersham, J. (2001) “Jane Jacob's Critique of Zoning: From Euclid to Portland and Beyond”. *Boston College Environmental Affairs Law Review 28 B. C. Env Rev. 552 (2000-2001).* Pg 548-552.

Winkelman, S., Allison B., and Chuck K. (2010) "Planning for economic and environmental resilience." Transportation Research Part A: Policy and Practice 44.8: 575-586.

Woelfle, M., Piero, O., and Matthew H.T. (2011) "Open science is a research accelerator." *Nature chemistry 3.10*: 745-748.

Zaltz Austwick, M., O'Brien, O., Strano, E., Viana, M. (2013) "The Structure of Spatial Networks and Communities in Bicycle Sharing Systems". *PLoS ONE* 8(9): e74685. doi:10.1371/journal.pone.0074685 Pg 11.

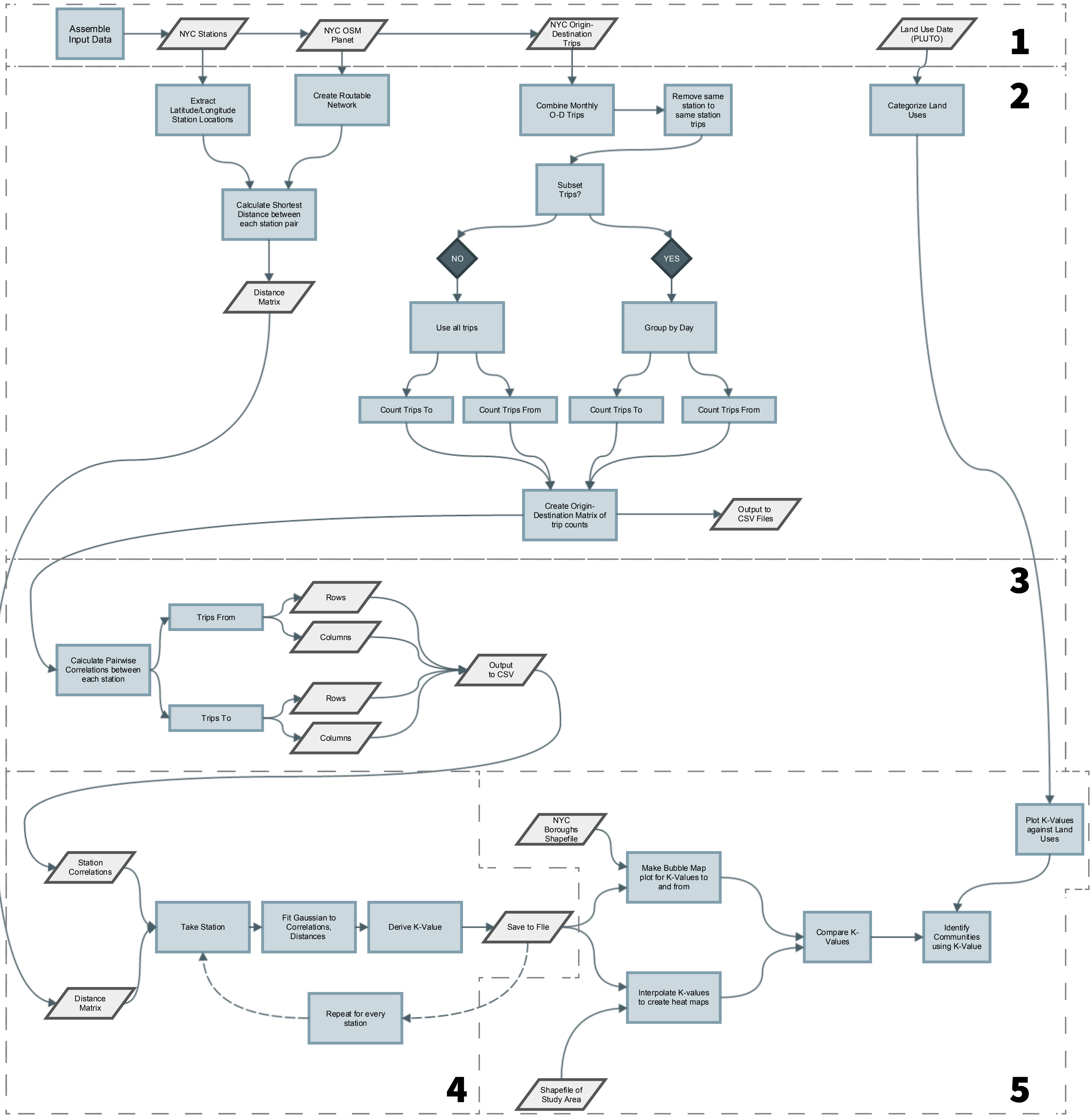
Appendices

Appendix 1 - Github Repository

The source code repository for this analysis can be accessed at the following link:

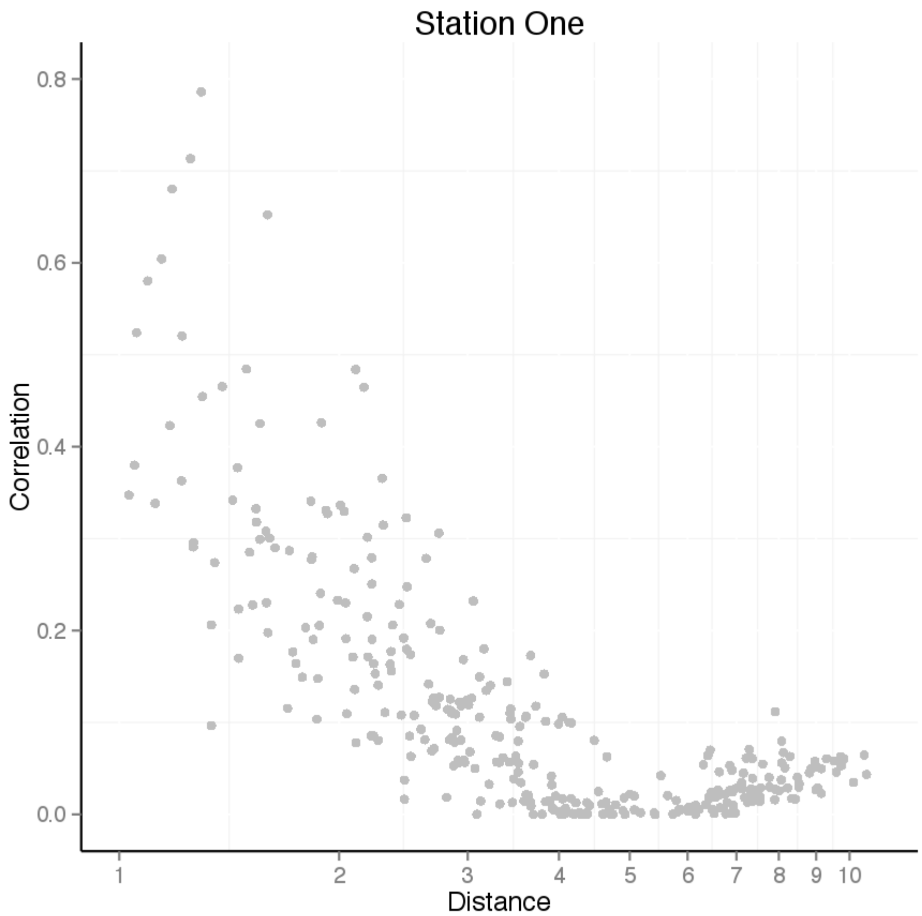
https://github.com/rustyb/bike-correlations/tree/master

Appendix 2
Procedural Diagram


1
Assemble Input Data
NYC Stations
NYC OSM Planet
NYC Origin-Destination Trips
Land Use Date (PLUTO)
2
Extract Latitude/Longitude Station Locations
Create Routable Network
Combine Monthly O-D Trips
Remove same station to same station trips
Categorize Land Uses
Calculate Shortest Distance between each station pair
Subset Trips?
NO
YES
Distance Matrix
Use all trips
Group by Day
Count Trips To
Count Trips From
Count Trips To
Count Trips From
Create Origin-Destination Matrix of trip counts
Output to CSV Files
3
Trips From
Rows
Columns
Calculate Pairwise Correlations between each station
Output to CSV
Trips To
Rows
Columns
Plot K-Values against Land Uses
NYC Boroughs Shapefile
Station Correlations
Make Bubble Map plot for K-Values to and from
Take Station
Fit Gaussian to Correlations, Distances
Derive K-Value
Save to File
Compare K-Values
Identify Communities using K-Value
Interpolate K-values to create heat maps
Distance Matrix
Repeat for every station
4
Shapefile of Study Area
5

Appendix 3
Large Graphs

Figure 2 - Example distribution of correlations and distances for Station One and Station Two.



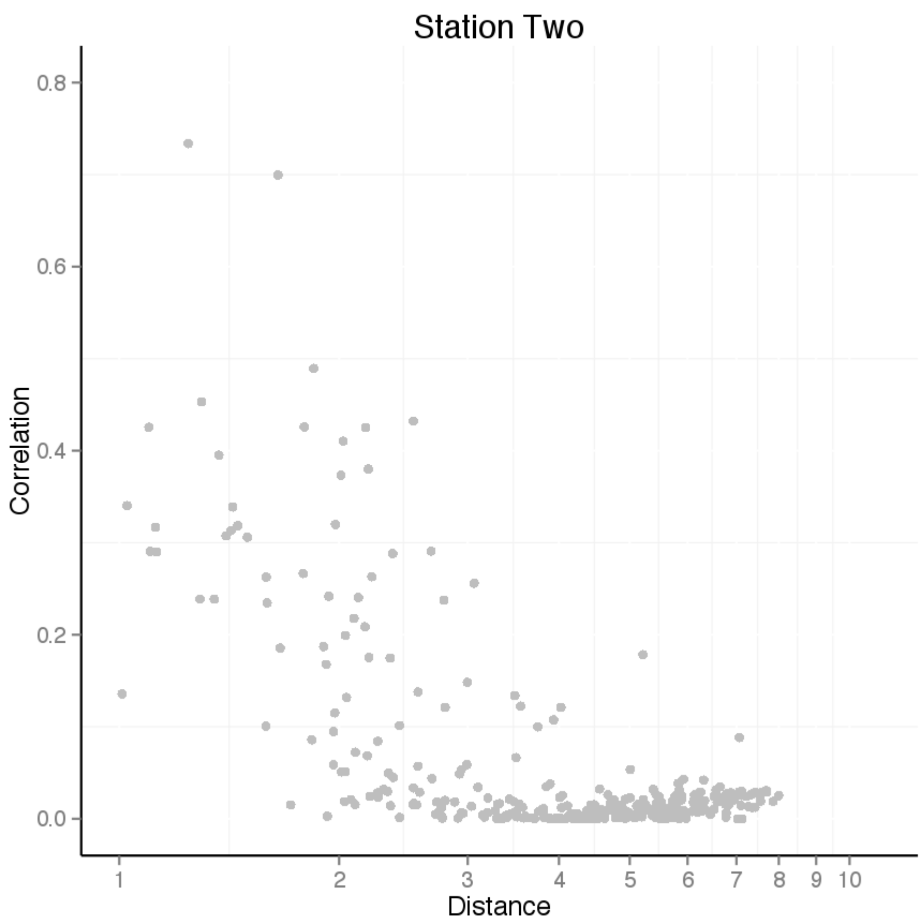


Figure 3 - Distance decay model examples.

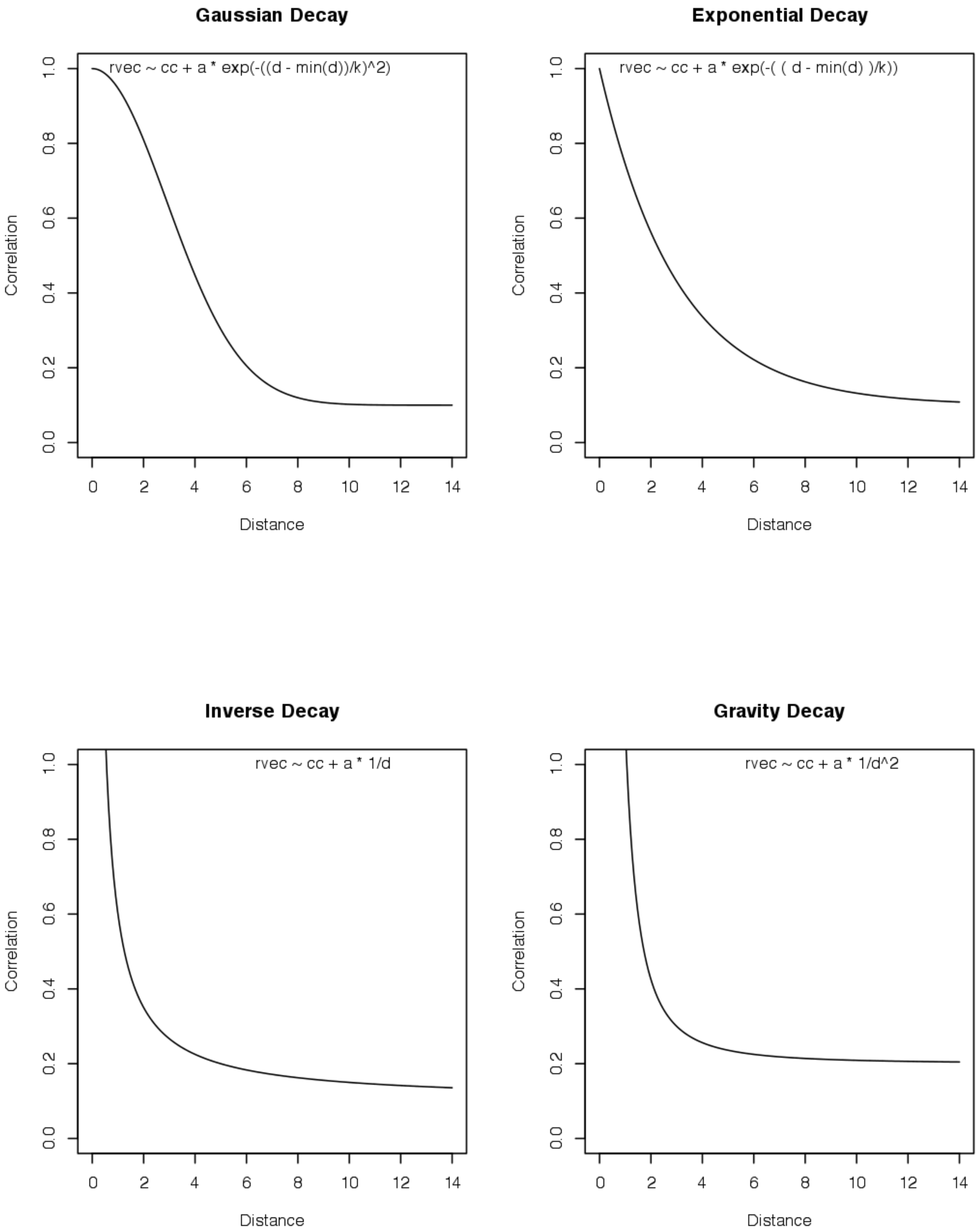


Figure 14 - Comparison of fitted distance decay models on two sample stations and resulting sum of squared residuals.

Station	S1	S2
Gaussian	1.78	1.71
Exponential	1.87	1.92
Inverse	4.69	3.95
Gravity	5.86	4.07

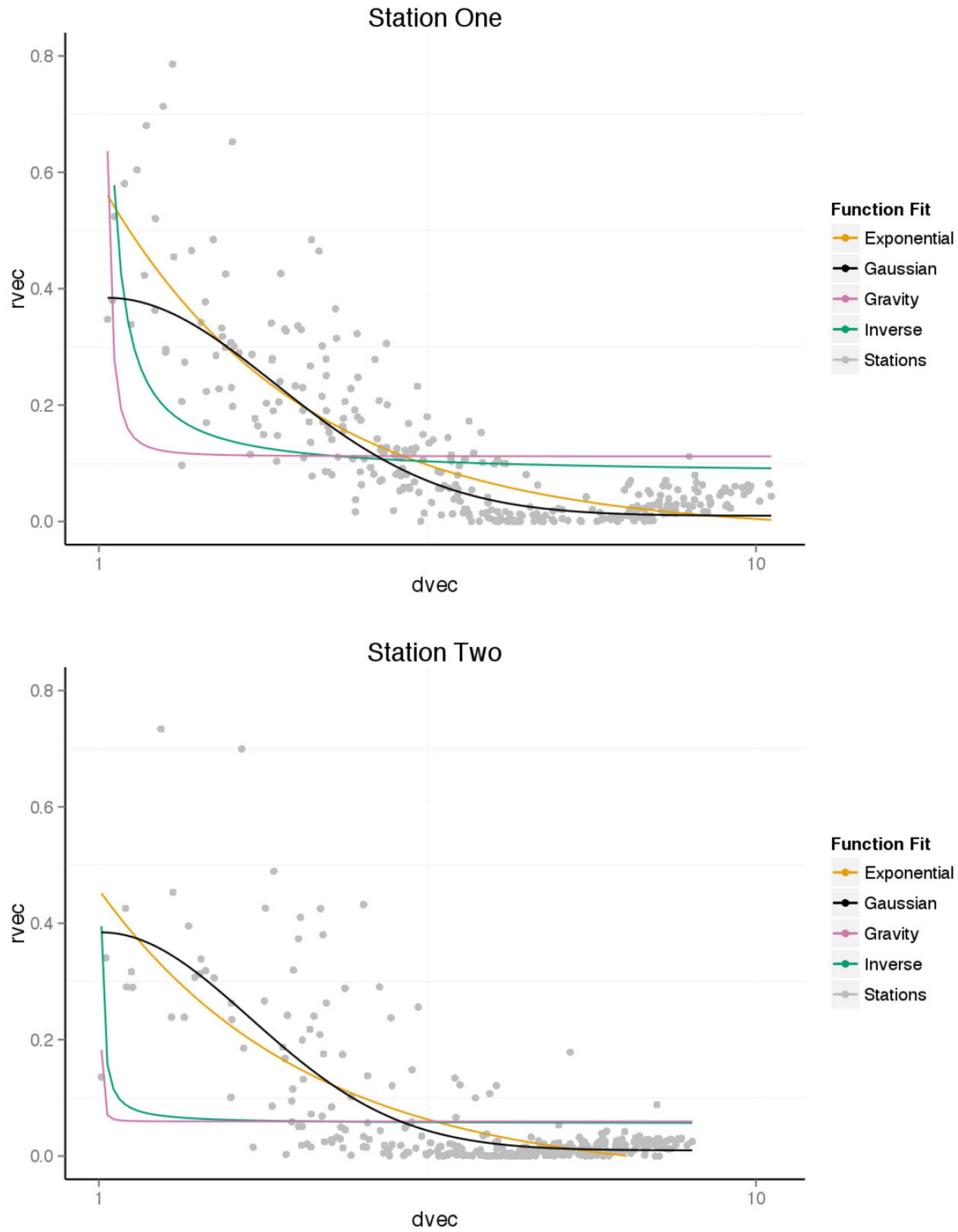
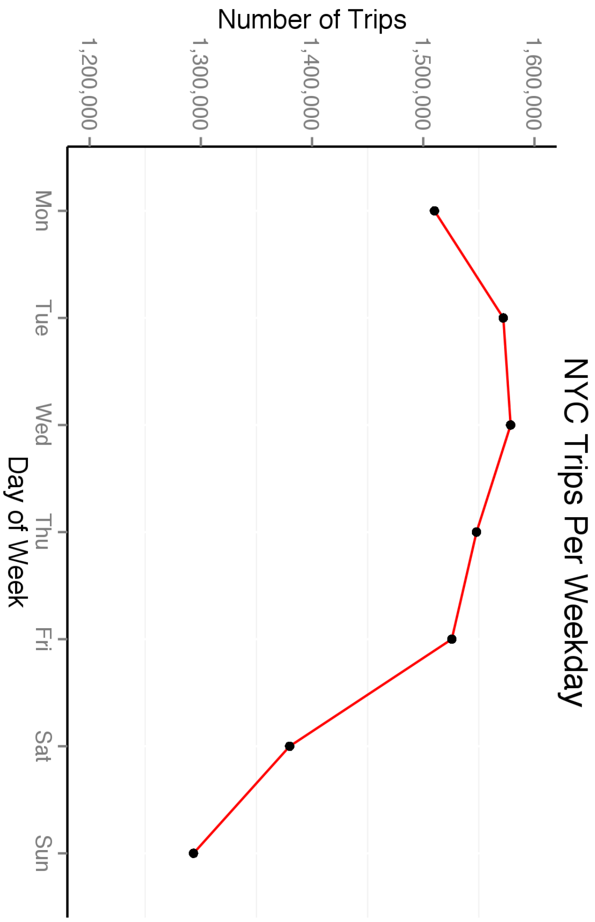


Figure 15 - NYC Trips From and Trips To K-Values

Figure 16 - Total Trips per weekday. Number of trips grouped by journey trip duration per day



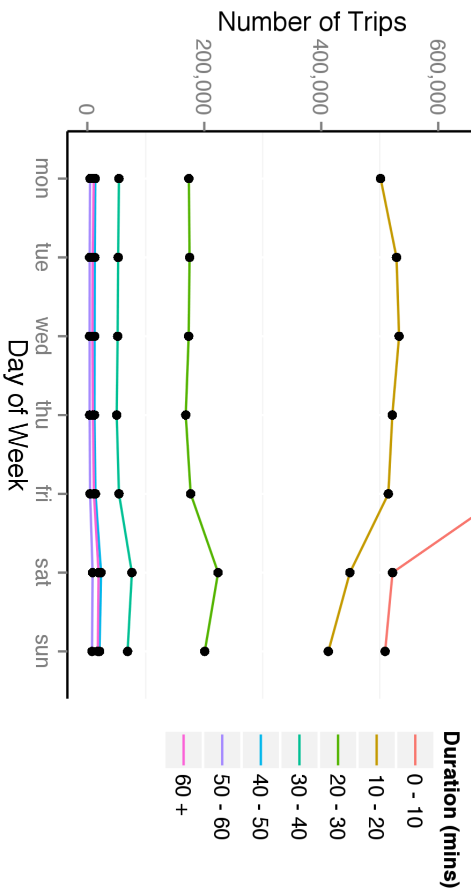


Figure 17 - NYC K-Values for the 90% and 100% quantiles.

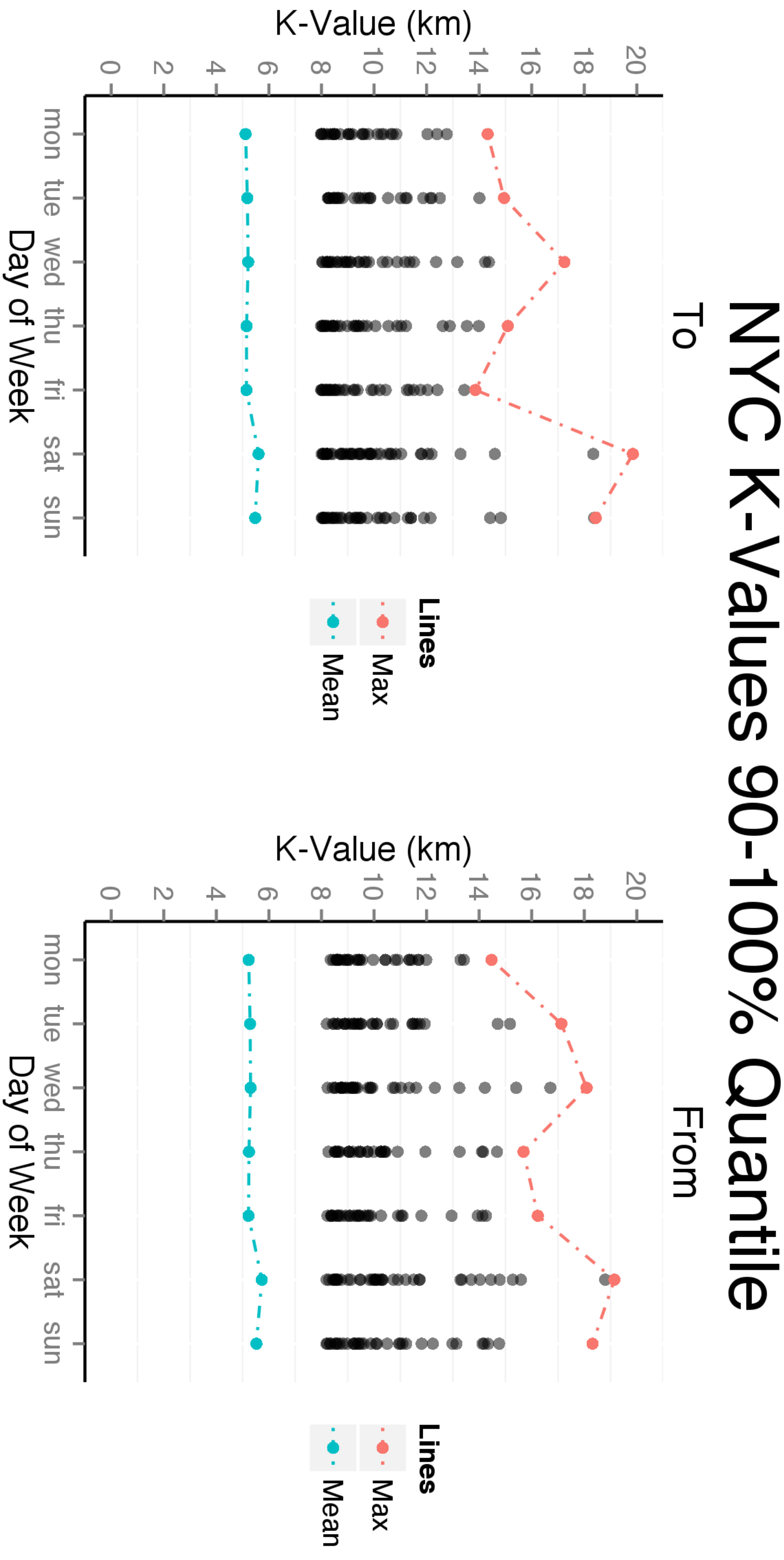
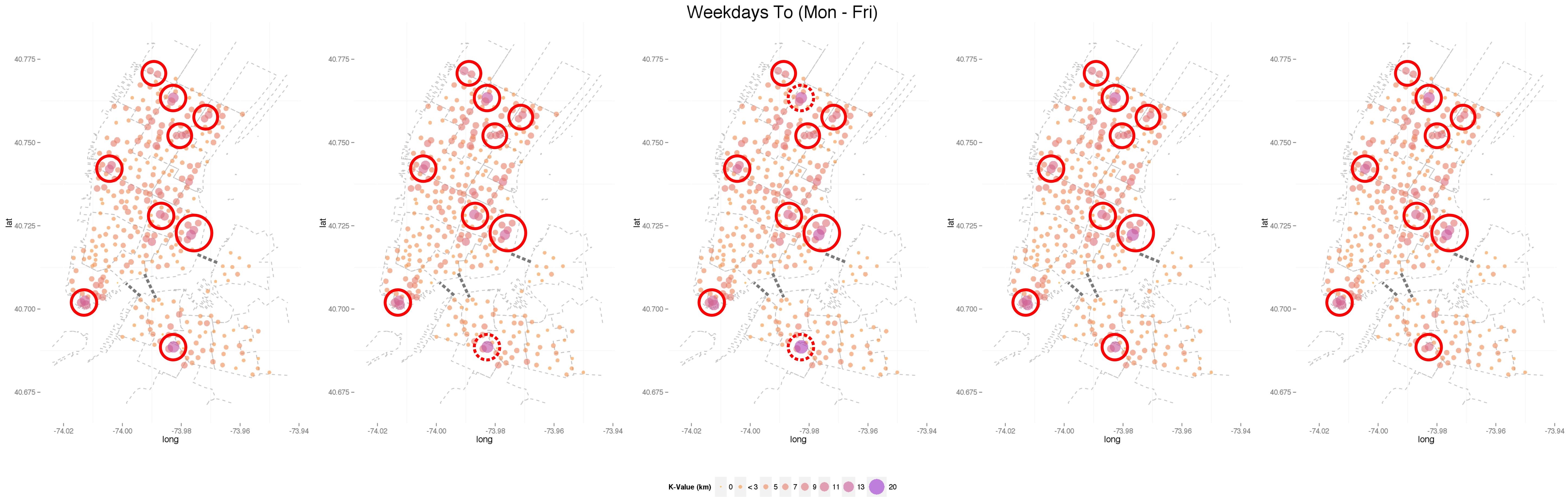


Figure 19 - New York K-Values on Weekdays (Monday - Friday) with high K-value communities identified



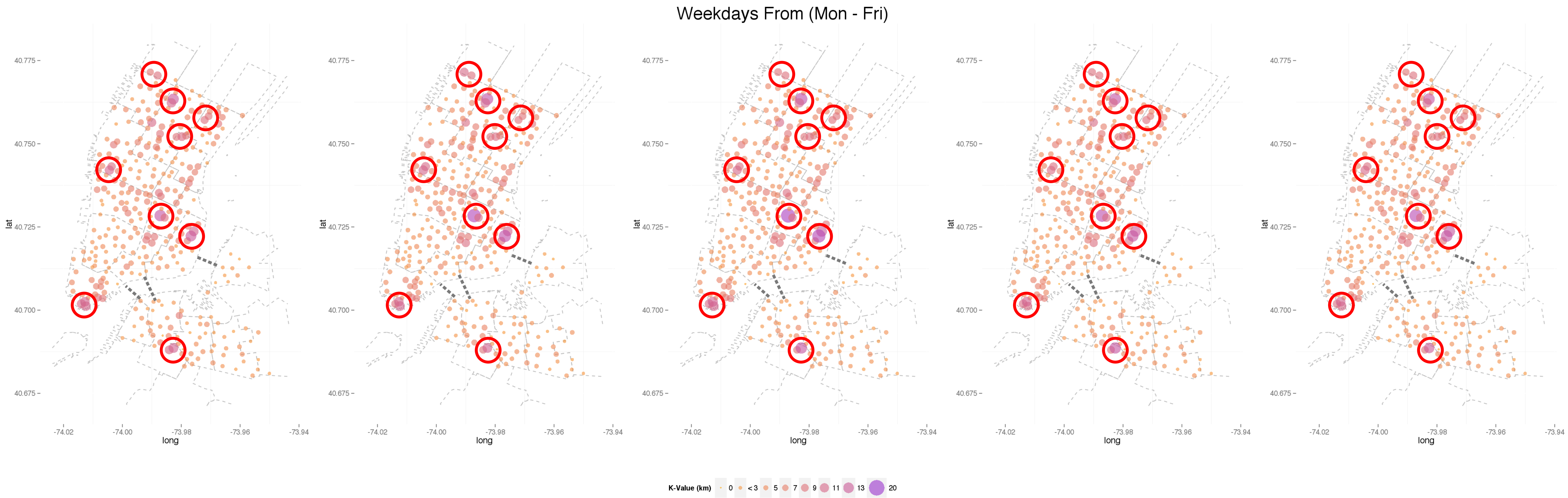
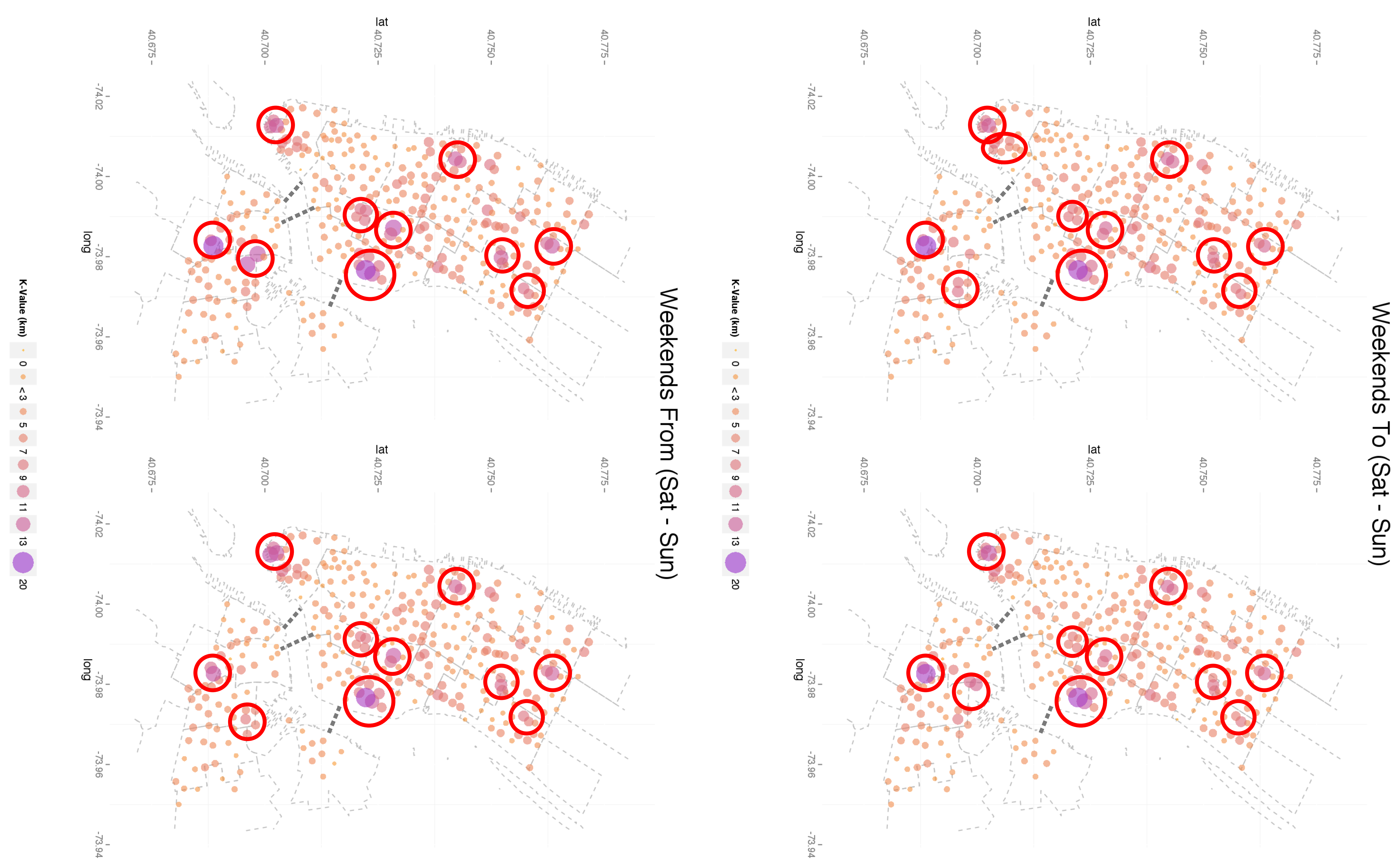
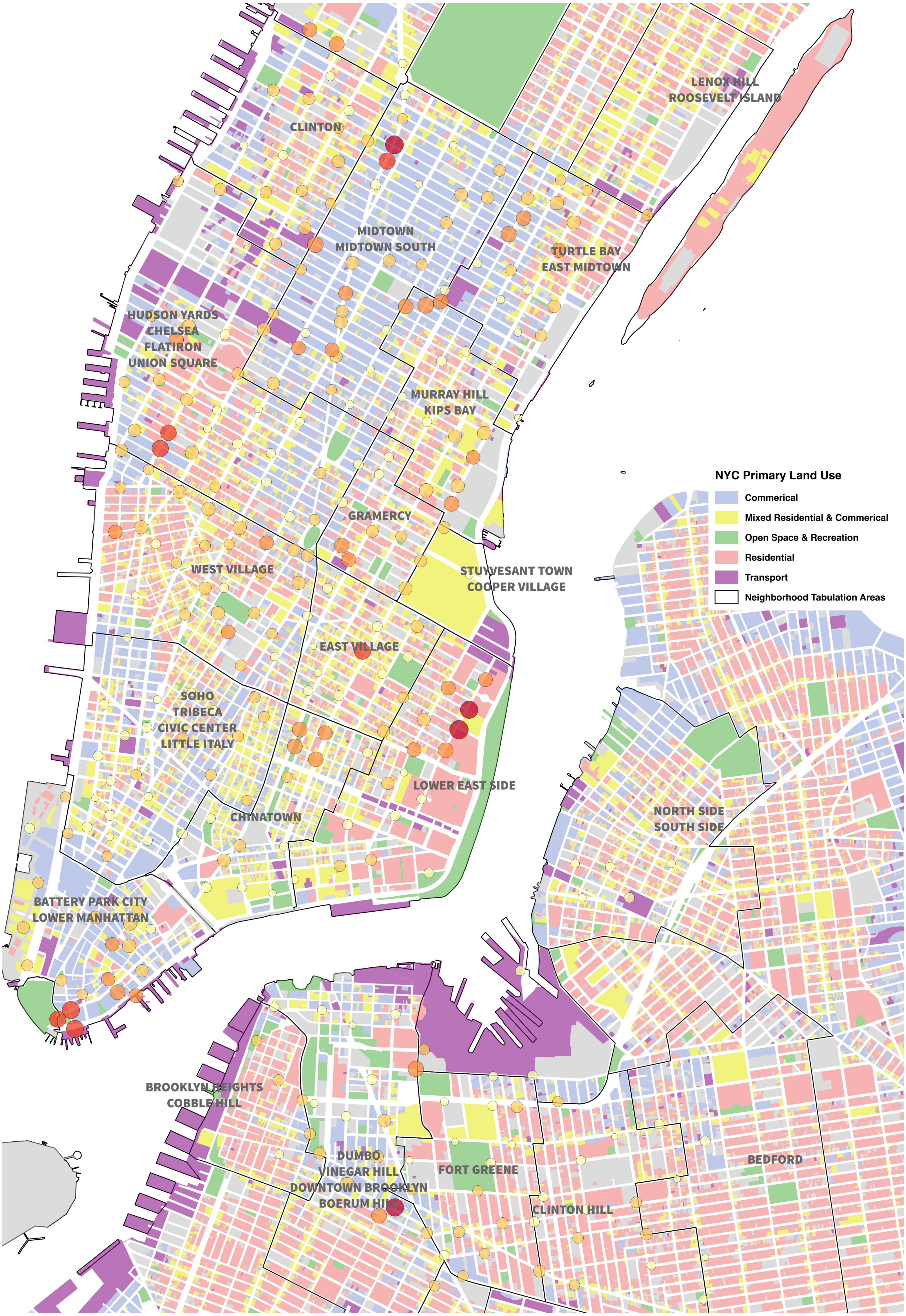


Figure 20 - New York K-Values on Weekends (Saturday - Sunday) with high K-value communities identified



Appendix 4
PLUTO Large Map


LENOX HILL
ROOSEVELT ISLAND
CLINTON
MIDTOWN
MIDTOWN SOUTH
TURTLE BAY
EAST MIDTOWN
HUDSON YARDS
CHELSEA
FLATIRON
UNION SQUARE
MURRAY HILL
KIPS BAY
GRAMERCY
WEST VILLAGE
STUYVESANT TOWN
COOPER VILLAGE
EAST VILLAGE
SOHO
TRIBECA
CIVIC CENTER
LITTLE ITALY
LOWER EAST SIDE
CHINATOWN
NORTH SIDE
SOUTH SIDE
BATTERY PARK CITY
LOWER MANHATTAN
BROOKLYN HEIGHTS
COBBLE HILL
DUMBO
VINEGAR HILL
DOWNTOWN BROOKLYN
BOERUM HILL
FORT GREENE
CLINTON HILL
BEDFORD
NYC Primary Land Use
Commerical
Mixed Residential & Commerical
Open Space & Recreation
Residential
Transport
Neighborhood Tabulation Areas